Issues in Using Self-Organizing Maps in Human Movement and Sport Science

Universitatea
Transilvania
din Brașov
FACULTATEA DE MATEMATICĂ
ȘI INFORMATICĂ

Ionescu Vlad,

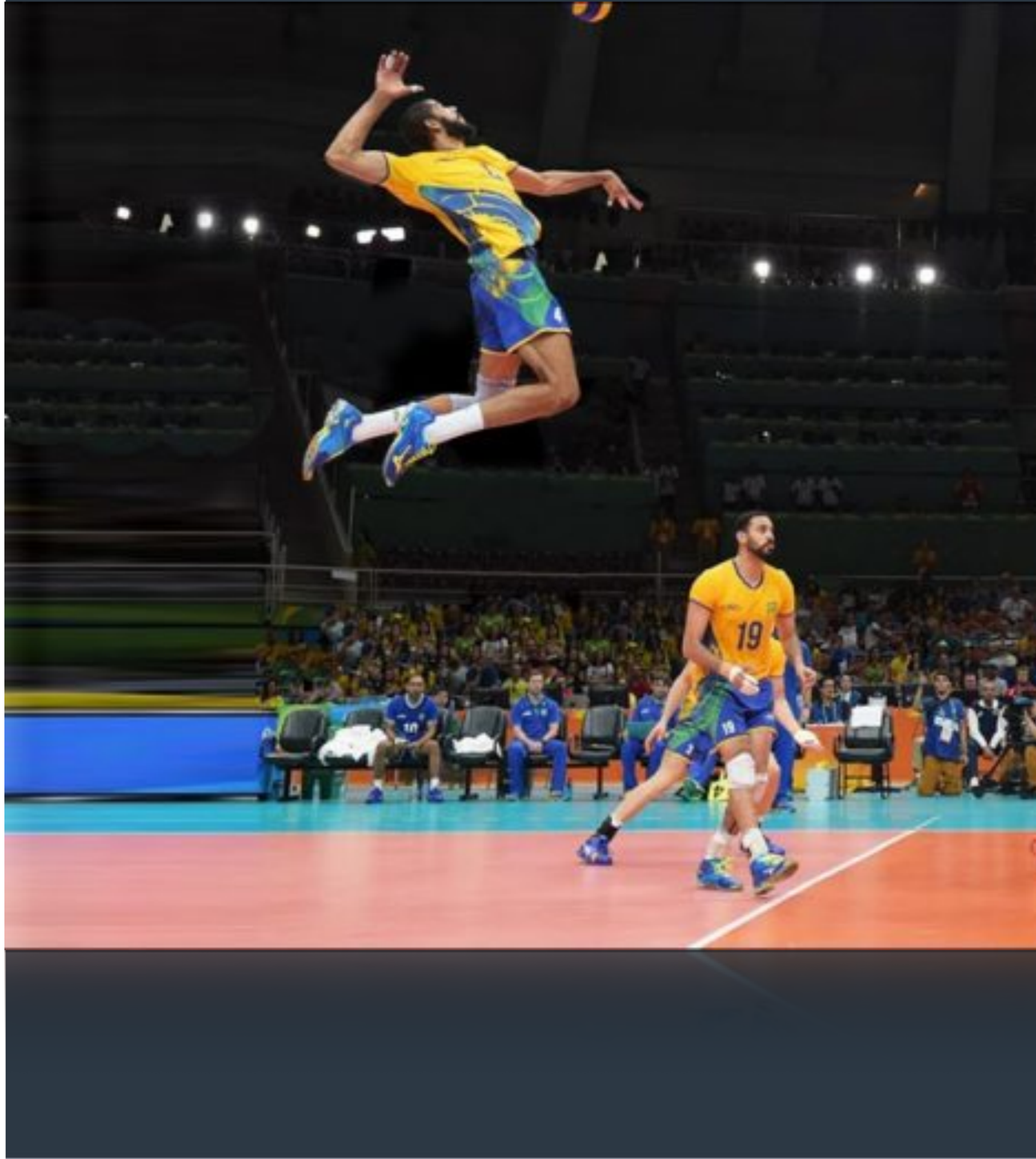
Transilvania University of Brașov

ROMANIA

Self-Organizing Maps

A **self-organizing map** (**SOM**) or **self-organizing feature map** (**SOFM**) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a **map**, and is therefore a method to do dimensionality reduction. Self-organizing maps differ from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as backpropagation with gradient descent), and in the sense that they use a neighborhood function to preserve the topological properties of the input space.

Self-Organizing Maps (SOMs) are steadily more integrated as data-analysis tools in human movement and sport science. One of the issues limiting researchers' confidence in their applications and conclusions concerns the (arbitrary) selection of training parameters, their effect on the quality of the SOM and the sensitivity of any subsequent analyses. In this paper, we demonstrate how quality and sensitivity may be examined to increase the validity of SOM-based data-analysis.



Notations

TE

Topographical error

The TE represents the percentage of input vectors for which the best-matching unit and second-best-matching unit are not neighbours (a measure of the continuity of the map)

QE

Quantization error

The QE represents the average Euclidean distance between a normalized input vector and its (trained) best-matching unit’s weight vector

CE

Combined error

The CE is the average Euclidean distance between an input vector and its second best-matching unit, passing first through the best-matching unit and then through the shortest path of neighbouring units towards the second best-matching one

Datasets

SOMs were used to calculate the coordination variability over ten trials. The research question concerned the difference in coordination variability between top level and junior players and male and female players (between subjects design). The previously published study (Serrien, Ooijen, et al., 2016b) found no significant interaction effect between gender and expertise on coordination variability, so for the present study, we concentrated only on the main effects (they found a significant main effect of gender, but not of expertise)

SOM analyses

The first step consisted of systematically adjusting the SOM parameters and investigating their effect on the quality of the SOM (CE, TE and QE) and on the effect size (ES gender) related to the research questions. To compare the different options of the SOM parameters, the dependent variables were tested with a non-parametric Friedman test and follow-up Wilcoxon tests for parameters with more than two options ($\alpha = 0.05$ with post-hoc Bonferroni corrections). To guide the decision- making process, we prioritized the options that produced significantly lower CE because it represents both map and accuracy

Table 1: SOM parameters and their options (see Appendix for a short explanation)

SOM parameters	**Options**
Lattice	Rectangular, Hexagonal
Initialization	Linear, Random [a]
Map Size	Small, Medium, Big
Training Length	Short, Normal, Long
Neighborhood function	Gaussian, cut-of Gaussian, bubble, Epanechikov
Training type	Sequential, Batch

[a] We ran five different simulations for the random weight vector initialization.

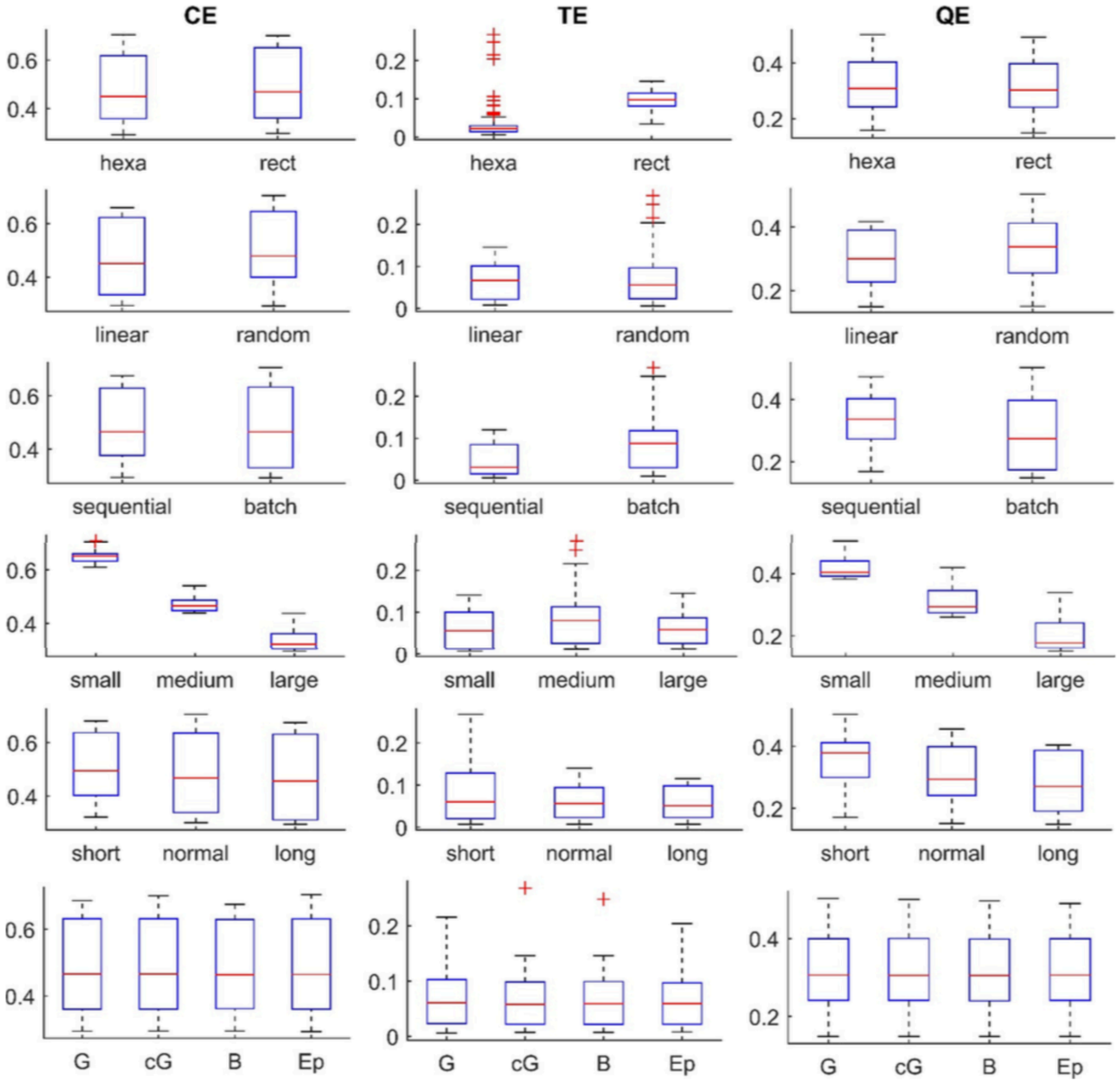


Figure 1: Dataset 1. Boxplots representing the CE (left), TE (middle) and QE (right) of the several simulations split between the options of the six SOM parameters: lattice (row 1), initialization (row 2, random is concatenated over all 5 seeds), training algorithm (row 3), map size (row 4), training length (row 5) and neighborhood function (row 6: G = Gaussian, cG = cut-off Gaussian, B = bubble, Ep = Epanechikov).

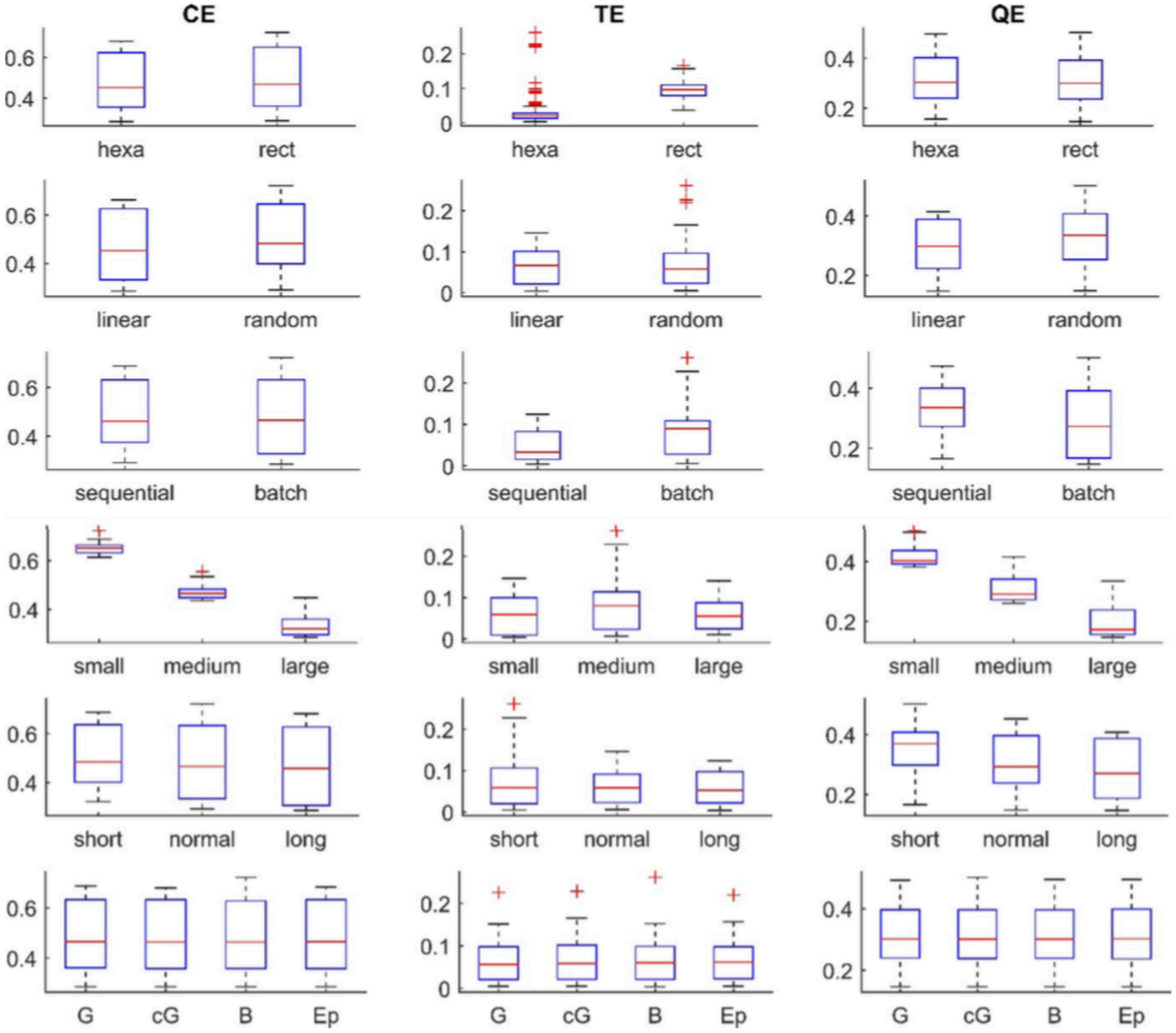


Figure 2: Dataset 2. Boxplots representing the CE (left), TE (middle) and QE (right) of the several simulations split between the options of the six SOM parameters: lattice (row 1), initialization (row 2, random is concatenated over all 5 seeds), training algorithm (row 3), map size (row 4), training length (row 5) and neighborhood function (row 6: G = Gaussian, cG = cut-off Gaussian, B = bubble, Ep = Epanechikov).

Sensitivity

The effect sizes related to the research questions of both datasets for all simulations are shown in Figure 3. The ES for the gender factor was significantly affected by SOM training length (short, normal < long). However, more importantly, all simulations demonstrated that this ES was in the same direction (male players exhibiting a smaller coordination variability than female players).

The ES for the within subjects factor year was significantly affected by training type (sequential > batch), map size (small > medium > large) and training length (short > normal > long). Again here as well, more importantly is that all simulations demonstrated that this ES was in the same direction (smaller coordination variability in year two). The (in)consistency in ES is an important validity check of the true (false) difference in a statistical test of coordination variability between groups/conditions.

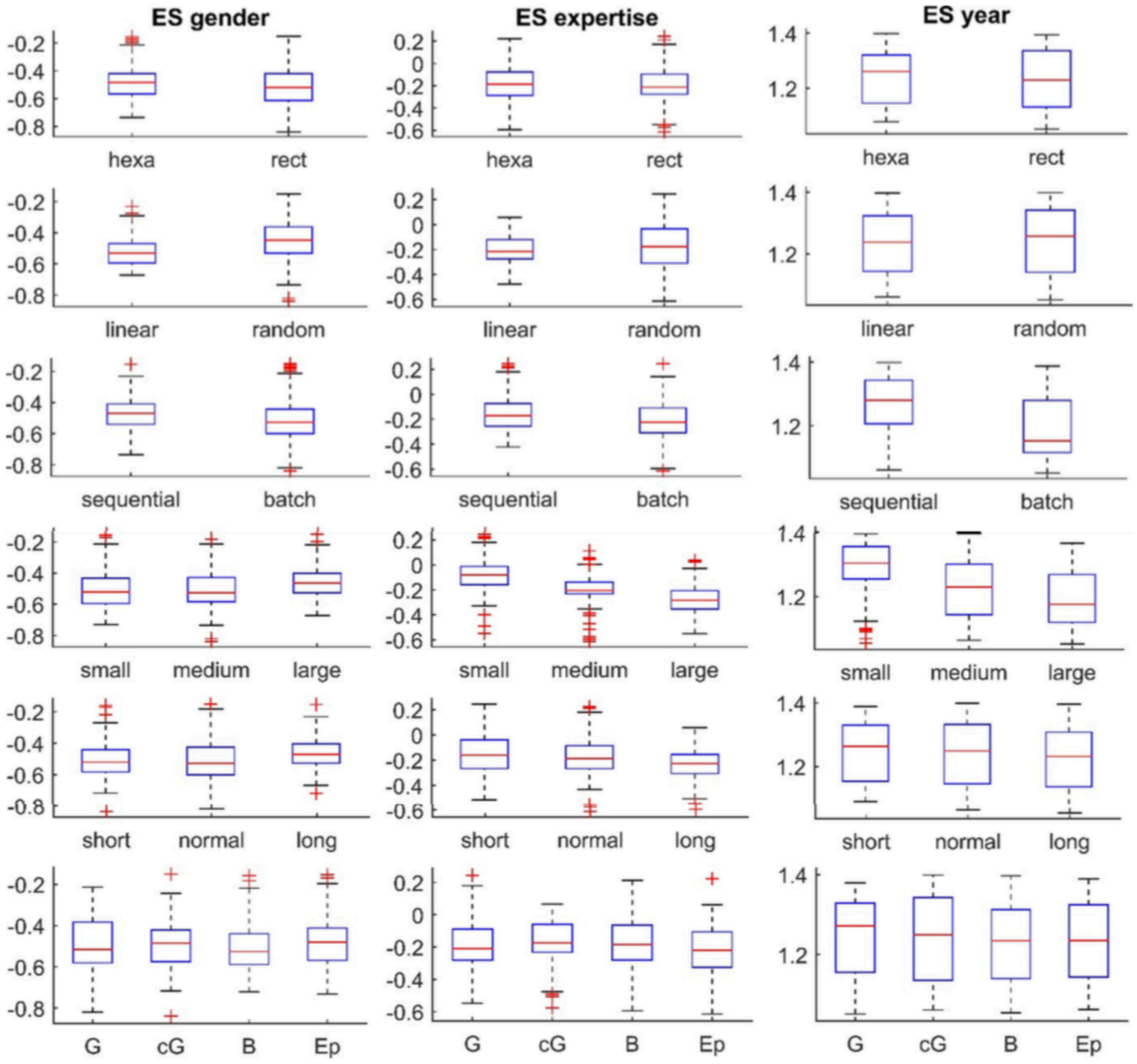


Figure 3: Boxplots representing the ES gender (left) and ES expertise (middle) for dataset 1 and the ES year for dataset 2 (right) of the several simulations split between the options of the six SOM parameters. lattice (row 1), initialization (row 2, random is concatenated over all 5 seeds), training algorithm (row 3), map size (row 4), training length (row 5) and neighborhood function (row 6: G = Gaussian, cG = cut-off Gaussian, B = bubble, Ep = Epanechikov). ES gender: negative values indicate greater variability for female players. ES expertise: negative values indicate greater variability for youth players. ES year: positive values indicate lower variability in the second year.

Step two

The results of training the SOM on a fewer number of trials (fewer input vectors) and variables (fewer vector components) are shown in Figures 4 and 5 for datasets 1 and 2 respectively. Every cell in the plots shows the median value of the errors across all players. Note that all rows can be compared to each other, but not all columns, only those columns with the same number of components. The numbers referring to the different variables are as follows: (1-3) pelvis sagital plane tilt, lateral tilt, rotation; (4-6) trunk sagital plane tilt, lateral tilt, rotation; (7-9) shoulder in/external rotation, ab/adduction, horizontal ab/adduction; (10) elbow flexion/extension. For both datasets, it is clearly visible that using fewer trials results in lower CE and QE, while the TE is not visually affected by reducing the number of trials. The larger CE and QE for the six right most columns is a trivial result, because the dimensionality is significantly lower in these simulations (CE and QE are distance measures). The SOM errors associated with the angular velocities were larger compared to the angles, probably because they showed more within subject variability

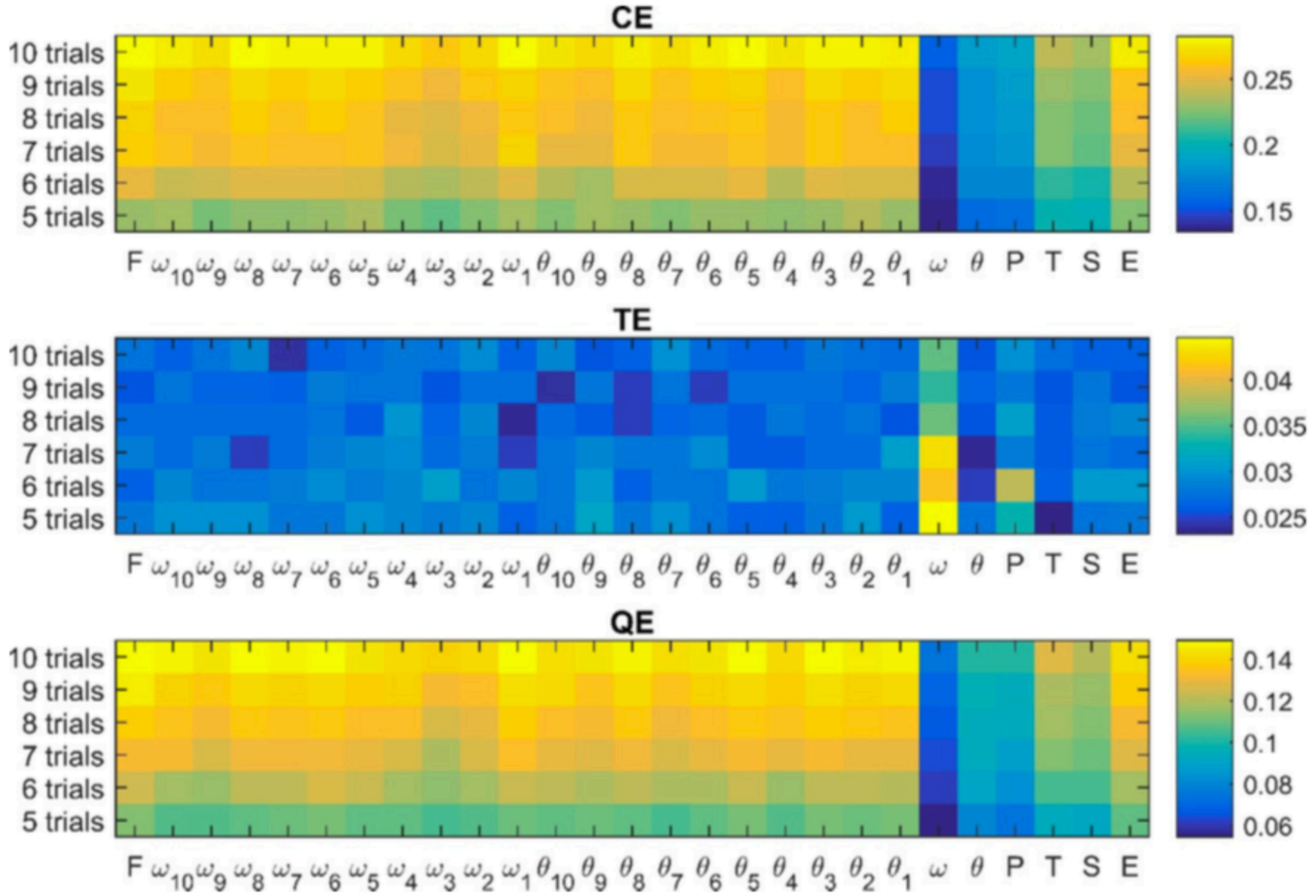


Figure 4: Dataset 1. Colour code plots representing the effect on the map quality of manipulating the number of trials used to train the SOM (rows) and the components representing the input vectors (columns). The first column (F) represents the full dataset (20 components). The other columns represent simulations where a specific angular velocity or angle time series was dropped from the dataset (ω_i or θ_i; see text for the numbers), where we dropped all angular velocities and angle time series (ω, θ), all pelvis variables (P), all trunk variables (T), all shoulder variables (S) or the elbow variables (E).

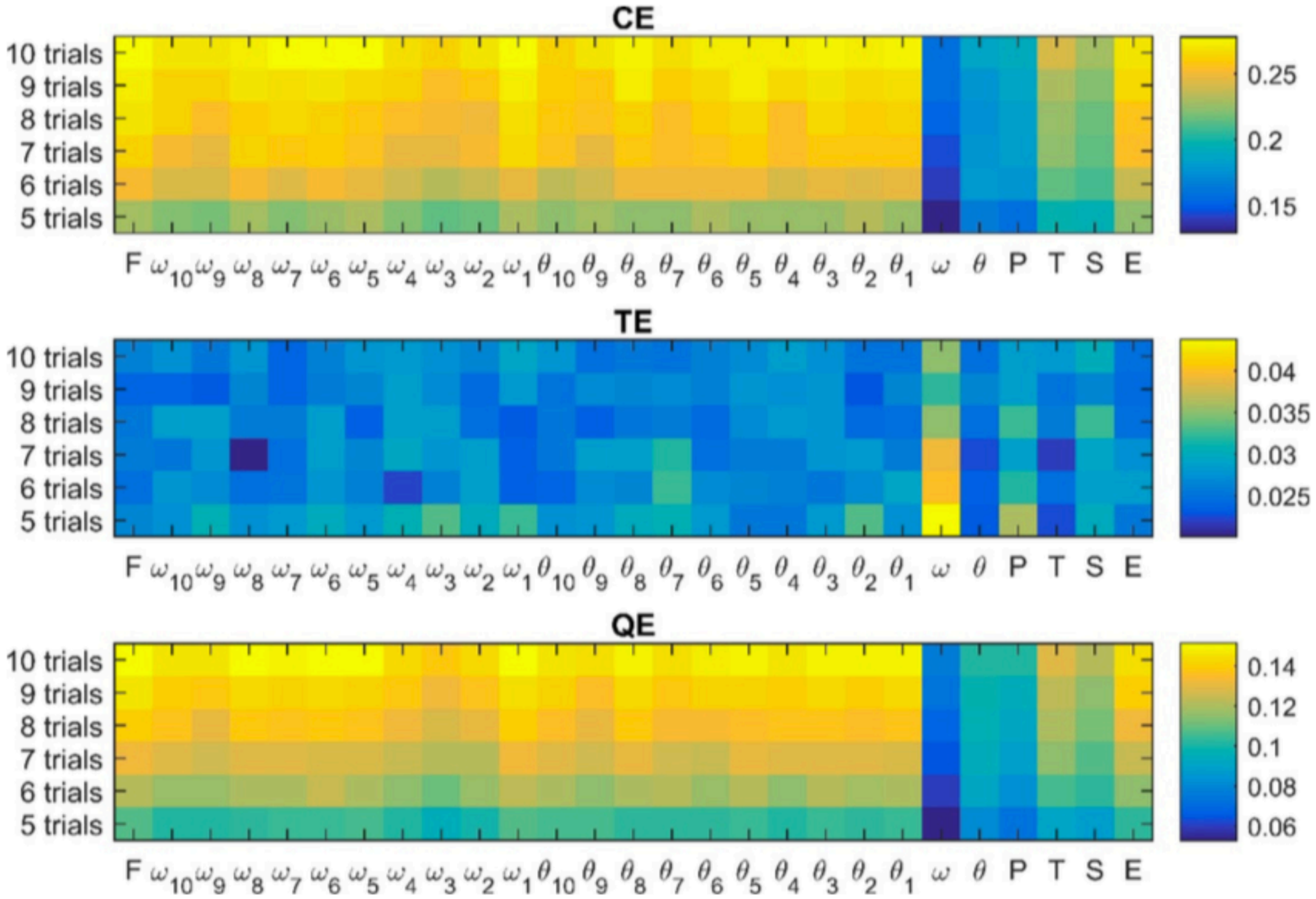


Figure 5: Dataset 2. Colour code plots representing the effect on the map quality of manipulating the number of trials used to train the SOM (rows) and the components representing the input vectors (columns). The first column (F) represents the full dataset (20 components). The other columns represent simulations where a specific angular velocity or angle time series was dropped from the dataset (ω_i or θ_i; see text for the numbers), where we dropped all angular velocities and angle time series (ω, θ), all pelvis variables (P), all trunk variables (T), all shoulder variables (S) or the elbow variables (E).

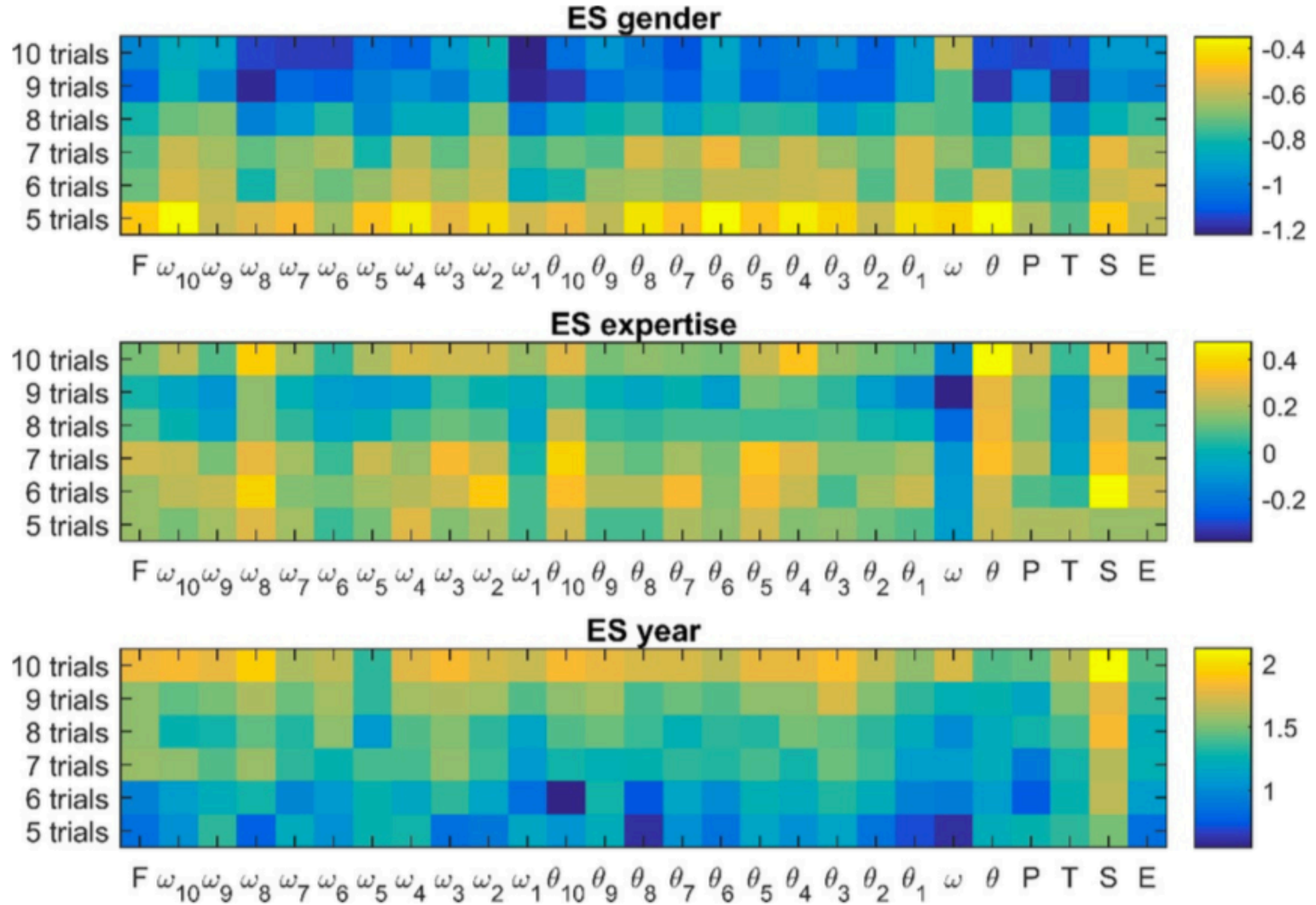


Figure 6: Colour code plots representing the sensitivity of the effect sizes on manipulating the number of trials used to train the SOM (rows) and the components representing the input vectors (columns). The first column (F) represents the full dataset (20 components). The other columns represent simulations where a specific angular velocity or angle time series was dropped from the dataset (ω_i or θ_i; see text for the numbers), where we dropped all angular velocities and angle time series (ω, θ), all pelvis variables (P), all trunk variables (T), all shoulder variables (S) or the elbow variables (E). (Top and middle plot: dataset 1, bottom plot: dataset 2).

Discussion and Conclusions

Whether the main conclusions in this study can be extrapolated to similar studies in human movement and sport science with time series data is unknown and a formulation of a set of guidelines for consistent use of SOM parameters and number of variables is not relevant. The better strategy would be to report this kind of analysis of the SOM quality and sensitivity of the subsequent analyses. For the datasets and corresponding research questions analysed in this article, we can robustly confirm our previous findings of lower coordination variability for male volleyball players and no difference between junior and elite players and the (unpublished) finding that the coordination variability in junior players was lower in the second year of the longitudinal study. In the present study, these SOM-based analyses were rather simple (between and within subjects differences), so we should examine also more complex analyses like clustering or applications where two or more SOMs are used in series