Quality & Quantity (2007) 41:93–104
DOI 10.1007/s11135-005-3148-0

Optimising a Probabilistic Model of the Development of Play in Soccer

J. CASTELLANO-PAULIS[1,*], A. HERNÁNDEZ-MENDO[2], VERÓNICA MORALES-SÁNCHEZ[2] and M. T. ANGUERA-ARGILAGA[3]
[1]*IVEF – Instituto Vasco de Educación Física, Universidad del País Vasco, Vitoria, Spain;* [2]*Universidad de Málaga, Málaga, Spain;* [3]*Universidad de Barcelona, Barcelona, Spain*

Abstract. We present a probabilistic model based on the one developed by Hernández Mendo and Anguera (Revista de Psicologíca Social, 16(1), 71–93, 2001). Here we have tried to break down the interaction contexts that the opposing teams are able to generate and transform during the game. We are aware that a given player or team does not produce consistent behaviour in similar situations. However, a degree of uncertainty is assumed to exist regarding whether the results obtained are a specific function of the analysis used. In order to carry out this research a category system which optimized that used in the previous model was developed. This system should enable the interaction between teams to be observed within the actual play of a soccer game. A lag sequential analysis was performed on the basis of a coding of the behavioural flow. After describing the behavioural patterns obtained a probabilistic model of the development of play in soccer is proposed.

Key words: behavioural patterns, interaction context, *lag-log* design, lag sequential analysis, Probabilistic model, soccer game

1. Introduction

In competitive soccer increasing importance is being attached to the rigorous and detailed observation of the strategies used by opposing teams with respect to both the ball in play and the stationary ball. Although soccer is by nature a varied game – of overlapping areas and quasi-complete and quasi-permanent disorder – certain aspects tend to follow one another during the game (Castellano, 2000). Furthermore, we believe that there is a greater degree of predictability when referring to the same team, for as Chappuis and Thomas (1988) argue there is a "*basic team personality*"; thus, one team in a team sport may appear dissimilar to another even though both have the same number of players. "*The basic personality of a*

*Author for correspondence: Julen Castellano-Paulis, Facultad de Ciencias de la Actividad Física y el Deporte, Carretera de Lasarte, s/n, 01007 Vitoria, Spain, E-mail: julen.castellano@ehu.es

team is the result of a cognitive and affective interaction between its members, one that takes the form of a particular style and which, in the physical and psychological sense, is shared by the majority of the players" (Chappuis and Thomas, 1988, p. 97).

Parlebas (1987, p. 123) supports this idea when he states that "*every sports player well knows that a team has its own style and its preferred 'savoir-jouer' (know-how regarding play), the players producing a range of preferential praxemes. A socio-motor microculture, with its tactical formulas and its motor 'paths,' develops within each team*". Knowing how the opposing team will approach the game, especially in a series of given situations, or knowing where one's own team-mates are going to position themselves in certain situations favours a more dynamic reading of the game. Furthermore, knowing the shared 'language' of soccer, the idiosyncratic aspects of one's team-mates, and being able to predict certain behaviours on the basis of others facilitates the 'obligatory' search for meaning in the motor behaviour engaged in by other players (whether team-mates or opponents) in order to make sense of the apparently disordered display of their actions.

Similarly, there are many other aspects, repeated time and again, which appear to be hidden within the game. These elements have a certain rhythm during the game, and the chronology of events – as well as their sequentiality – is highly important in this regard. A sequential analysis is a particular type of probabilistic process where each one of the behavioural events in a chain is dependent on both the initial event (criterion behaviour) and previous events (Gorospe, 1999 Unpublished doctoral thesis). The authors consider the competitive sport of soccer as a sequential process in which the next behavioural event is determined by the initial behaviour (criterion category) – provided they are considered as positive lags – and the antecedent event. The sequentiality of the events assumes independence and considers the probability of the transition of one event to another as being dependent upon the information contained in the present event – and, moreover, that it doesn't change over time, as the probabilities of transition between two events may not be different at two points in time (Hernández and Anguera, 2001).

Research into behavioural chains in soccer (Hernández, 1996, Unpublished doctoral thesis; Ardá, 1998, Unpublished doctoral thesis; Castellano and Hernández, 1999; Castellano, 2000, Unpublished doctoral Thesis) bring greater meaning to the game and, consequently, are more useful for managers or trainers. Being able to identify the paths of interaction between two teams enables us to determine which of these paths provide a team with the most effective mode of attack or defence in the game. In sum, it enables us to estimate those behavioural chains which occur with greater probability than would be predicted by chance.

In contrast to the model proposed by McGarry and Franks (1996), which is developed on the basis of a Markovian sequential analysis, the model presented here follows the proposal of Hernández and Anguera (2001), that is, it is based on a prospective lag sequential analysis. The difference between the methods of the two models resides in the fact that whereas in Markov chains only dependence on the previous behaviour is produced, a prospective lag analysis is able to predict, in an interpretable way, several lags (Gorospe, 1999, Unpublished doctoral thesis).

The first objective of this study was to optimise the model formulated by Hernández and Anguera (2001), complementing it with other studies carried out on behavioural 'maps' in soccer. This should enable the model's efficiency to be optimised.

2. Method

The present study continues the innovative line of research begun a few years ago in applying observational methodology to team sport contexts (Hernández, 1996; Ardá, 1998; Gorospe, 1999, Unpublished doctoral thesis). A new development here is that the proposed observational design – synchronic/diachronic design or *lag-log* – is situated within quadrant IV of those proposed by Anguera (1990). Both a lag sequential analysis and a generalisability analysis, applied to the quality of the data, were performed. This enabled application to a social context, namely, the psychosociology of sport. Finally, a way of optimising Hernández and Anguera's (2001) first probabilistic model of team behaviour is proposed.

2.1. STRUCTURE OF THE CATEGORY SYSTEM

The *ad hoc* observation tool used is based on the combination of category systems and field formats, and is termed SOCCAF (Castellano, 2000, Unpublished doctoral thesis). The configuration of the category systems of each of the criteria is exhaustive and mutually exclusive (E/ME). It was developed by combining all the behaviours of several dimensions for each criterion, such that, by forcing unidimensionality, an E/ME category system was produced – in which all the possible behaviours of the situation to be studied are included, without new ones being able to be incorporated. We believe that both approaches are necessary in order to build the framework that will enable the contextualised and exhaustive recording of the behavioural aspects of interaction considered relevant in the game.

For reasons of space, of the two criteria making up the field formats of play in soccer (SOCCAF) the present study only considers the criterion of interaction. After having grouped the 68 categories which go to make up this criterion in different ways according to the interaction contexts and

strategic behaviour that the teams may develop during the match, both when the ball is in play and stationary, a total of 29 categories that exhaustively and exclusively represent the dynamics of the game's interaction were used (Table I).

Table I. Categories of the system for observing play in soccer

	Start of possession
RMT	The defensive area of the observed team recovers the ball, with the midfield and defensive area of the opposing team in front of it
RAT	The defensive area of the observed team recovers the ball, with the whole of the opposing team in front of it
MRT	The midfield of the observed team recovers the ball, with the defensive area of the opposing team in front of it
MMT	The midfield of the observed team recovers the ball, with the midfield and defensive area of the opposing team in front of it
MAT	The defensive area of the observed team recovers the ball, with the whole of the opposing team in front of it
ART	The forward area of the observed team recovers the ball, with the defensive area of the opposing team in front of it
AMT	The forward area of the observed team recovers the ball, with the midfield and defensive area of the opposing team in front of it
GT	The goalkeeper of the observed team recovers the ball
IRFP	Goal kick, corner, penalty and free kick in favour of the observed team
IRFM	Throw-in in favour of the observed team
	Development of possession
RMC	The observed team maintains possession in the defensive area by passing the ball, with the midfield and defensive area of the opposing team in front of it
RAC	The observed team maintains possession in the defensive area by passing the ball, with the whole of the opposing team in front of it
ERC	The observed team maintains possession in the external area by passing the ball, with the defensive area of the opposing team in front of it
MRC	The observed team maintains possession in midfield by passing the ball, with the defensive area of the opposing team in front of it
MMC	The observed team maintains possession in midfield by passing the ball, with the midfield and defensive area of the opposing team in front of it
MAC	The observed team maintains possession in midfield by passing the ball, with the whole of the opposing team in front of it
ARC	The observed team maintains possession in the forward area by passing the ball, with the defensive area of the opposing team in front of it

Table I. Continued

AMC	The observed team maintains possession in the forward area by passing the ball, with the midfield and defensive area of the opposing team in front of it
AOC	The observed team maintains possession in the forward area by passing the ball, with all but the goalkeeper of the opposing team in front of it
TIR	The observed team shoots at goal
INT	The opposing team intercepts the ball
	End of possession
RMP	The midfield of the opposing team recovers the ball, with the defensive area of the observed team in front of it
RAP	The forward area of the opposing team recovers the ball, with the defensive area of the observed team in front of it
MRP	The defensive area of the opposing team recovers the ball, with the midfield and defensive area of the observed team in front of it
MMP	The midfield of the opposing team recovers the ball, with the midfield and defensive area of the observed team in front of it
MAP	The forward area of the opposing team recovers the ball, with the midfield and defensive area of the observed team in front of it
ARP	The defensive area of the opposing team recovers the ball, with the whole of the observed team in front of it
AMP	The midfield of the opposing team recovers the ball, with the whole of the observed team in front of it
PG	The goalkeeper of the opposing team recovers the ball

2.2. SUBJECTS

In order to perform the analysis for this study ten matches played during the second (knock-out) stage of the 1998 World Cup in France were coded and recorded. These matches were: France–Brazil; Croatia–Holland; France–Croatia; Brazil–Holland; France–Italy; Brazil–Denmark; Croatia–Germany; Holland–Argentina; France–Paraguay; and Brazil–Chile. The choice of number of matches to observe was made by optimising a measurement plan in a generalizability analysis.

2.3. MATERIAL

The material used for coding the behavioural flow in order to perform the statistical and sequential analyses was as follows: VHS video recorder, television monitor, laptop computer with Pentium I processor (166 MHz, 32 Mb RAM and 4100 Mb hard disk), SPSS v. 6.1 for *Windows*, SDIS-GSEQ v. 2.0 sequential analysis program (Bakeman and Quera, 1996) and the GT v. 1.0 program (Ysewijn, 1996). Once each match

had been recorded the data were filtered through an instructions file (*.gsq), developed by the authors, that was able to detect both formal and conceptual errors in the recording of the behavioural flow.

2.4. PROCEDURE

An observation protocol was designed in order to ensure that the observed data obtained from different groups of previously-trained observers were of good quality. Generalizability theory was used to estimate, among other things, the minimum number of matches required to enable the results to be generalized (11 matches).

In order to carry out the sequential analyses and determine the corresponding *max lag*, the same team of observers coded ten soccer matches from the second stage of the 1998 World Cup in France. The SDIS-GSEQ program (Bakeman and Quera, 1996) was subsequently used to detect the corresponding patterns.

3. Results

3.1. ANALYSIS OF DATA QUALITY

In order to compare data quality a soccer match from the 1998 World Cup in France was coded at three different points in time by two groups of observers, previously-trained in accordance with agreed criteria (Anguera, 1990). Agreed concordance was used in each of the two groups (Anguera, 1990). One of the two groups of observers coded the match twice. Once the coding was complete the intra- and inter-observer concordance was calculated using the Kappa index, generalisability theory and a study of correlations. The estimated indicators for the quality of the recorded data are shown in Table II (Castellano et al., 2000).

3.2. SEQUENTIAL ANALYSIS

The measures of sequentiality enable dependency relationships in the flow of behaviour produced by one or more players at the same time to be established. This type of analysis seeks to identify the probability of transition between behaviours that is greater than that predicted by chance (Sackett, 1987). This probability does not imply direct linear relationships between two consecutive events in time. Indeed, the relationship should not be regarded from a deterministic point of view, but rather in terms of probability or stochastically; that is, the first event is simply the antecedent and the second the consequent, there being a certain degree of probability of transition of an associative nature.

Table II. Concordance (inter- and intra) indices for the different macrocategories of the taxonomic system and the set of categories as a whole

Correlation coefficients		
Coefficient for the whole session	Inter-concordance	Intra-concordance
Pearson's correlation	0.99	0.99
Kendall's tau	0.89	0.94
Spearman's coefficient	0.96	0.98
Cohen's kappa indices		
Category group	Inter-concordance	Intra-concordance
Rule-based interruption categories	1.00	1.00
Start of possession categories	0.75	0.85
Continuation of possession categories	0.82	0.92
End of possession categories	0.84	0.95
Continuation of non-possession categories	0.83	0.91
Other categories	0.91	0.97
General concordance of the session with commission errors	0.88	0.94
General concordance of the session with commission and omission errors	0.74	0.86
Generalizability analysis		
(Two-facet design: categories and observers [C/O])		
Variance components		G. coefficient
Observers	Categories	Residual
0%	99%	1%	0.99

The SDIS-GSEQ program (Bakeman and Quera, 1996) was used to carry out the prospective sequential analysis. The directionality of the sequential transitions has a certain logic within the match play, such that there will be categories which are only analysed from either a prospective or retrospective point of view. The following table summarises all the patterns found (Castellano and Hernández, 1999; Castellano, 2000).

4. Discussion

The data obtained in the present study clearly show that both the observational tool developed and the training of observers met the planned objectives and passed the quality control of recorded data. We believe, therefore, that the study meets the methodological requirements of research.

Table III. Description of estimated behavioural patterns, showing: criterion behaviour, the categories involved in each pattern, its *max lag*, the number of links it has, and the type of bifurcation it contains

SUMMARY TABLE OF THE SEQUENTIAL ANALYSIS

Macrocategory	Type of analysis	Criterion behaviour	Categories of the pattern	*Max Lag*	Number of links	Type of bifurcation
Start of possession from a stationary ball	Prospective	IRFM	MMC, INT	INT	2	Linear
		IRFP	INT, MMC	MMC	2	Linear
Start of ball possession after recovery	Prospective	RA-T	RAC, MMC	RAC	3	Dyadic
		RM-T	MMC	MMC	2	Linear
		MA-T	–	–	–	–
		MM-T	ARC, INT	ARC	2	Dyadic
		MR-T	–	–	–	–
		G-T	RAC	RAC	1	Linear
		AM-T	–	–	–	–
		AR-T	ARC	ARC	3	Linear
Continuation of ball possession	Prospective	RM-C	MMC	MMC	2	Linear
		RA-C	RAC, MMC	RAC	5	Octadic
		ER-C	ARC, INT	INT	3	Dyadic
		MR-C	ARC, INT	ARC	3	Dyadic
		MM-C	ARC, INT, MMC	MMC	2	Dyadic
		MA-C	MMC, RAC	MMC	3	Linear
		AR-C	INT, ARC	INT	3	Dyadic
		AM-C	MMC	MMC	1	Linear
	Retrospective	AO-C	–	–	–	–
Loss of ball possession	Retrospective	RM-P	–	–	–	–
		RA-P	INT	INT	−1	Linear
		MR-P	–	–	–	–
		MM-P	MMC, RAC INT	RAC INT	−2	Triadic
		MA-P	–	–	–	–
		AR-P	ARC	ARC	−1	Linear
		AM-P	–	–	–	–
		P-G	ARC	ARC	−3	Linear
Complementary behaviours	Retrospective	TIR	ARC	ARC	−1	Linear
	Prospective	INT	MMC, PER	PER	1	Linear

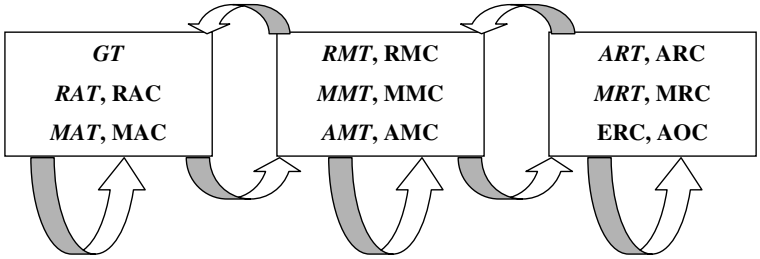

Figure 1. Probabilities of transition between the different categories of recovery and continuation of ball possession. The length of the arrow is related to the probabilities of transition.

It is demonstrated that those categories which imply a continuation in ball possession contain more extensive patterns, the majority of them containing significant excitatory transitions up to lag three. Approximately 75% of estimated patterns contain two or more lags (15 of 21).

The results obtained in the lag sequential analysis reveal aspects of transition that are crucial for the game. The behaviour of the teams with respect to the interaction contexts where it occurs produces patterns in which the transitions between the categories have situational or contextual 'proximity'. In other words, interaction contexts with a marked defensive component have a greater probability of transition toward contexts with a similar defensive component; likewise, contexts with a midfield or attacking component are more likely to appear as related to others of the same kind. Figure 1 shows the transition relationships between the different categories. The behaviours shown within the same square have a greater degree of activation among each other; this activation is present, but decreases, between adjacent squares and in both directions.

Noteworthy within this figure is an aspect which is crucial for the game, and thus the behaviours are shown in italics: categories with the same interaction contexts have a greater probability of transition or progression toward more attacking interaction contexts when their implicit strategic behaviour is ball recovery rather than continuation or maintenance of possession.

The diachronic evolution of transformation and/or transfer of the interaction contexts produced in the game, through the continuous movement of players and the ball, enables identification of: (1) which interaction contexts are most advisable for the game, guiding the actions of players and teams; and (2) the preferred moment for transformation and/or transfer of interaction patterns. For example, knowing when the defence should move forward or back a few metres, etc. In addition to these two possibilities, new ways of understanding in more detail how soccer play develops (Castellano, 2000, Unpublished doctoral Thesis) can be

incorporated, such as: (3) the preferred way of producing transformation and/or transfer of interaction contexts – by passing the ball, by movement of players or both these at the same time. For example, knowing how, in accordance with the state of play, to decide whether an individual action or a pass to a team-mate would improve the ball's position within the interaction, etc. Another option is: (4) identifying where or towards where is the best place for a transformation and/or transfer of a given interaction pattern to occur, for example, where a certain interaction context should be transferred – in midfield, close to the opposing team's area, etc.

On the basis of the results obtained we propose the following explanatory model which, optimising that described by Hernández and Anguera (2001), functions as a descriptive and predictive model of interaction in soccer.

As is shown in Figure 2 the transitions between interaction contexts are related through interaction proximity, that is, contexts which are implicitly more defensive (those where, in front of the ball, there is a high number of players from the opposing team, and from the observed one) tend to move toward contexts with the same defensive or midfield characteristics. Likewise, more attacking contexts (those where, in front of the ball, there is a smaller number of players from the opposing team and, to a certain extent, from the observed one) tend to flow toward contexts of a similar kind. It can also be seen that the most sought after contexts are RA, MM and AR, allowing greater knowledge of transitions between contexts. Among the main behaviours which appear in the game we include interception (INT), carried out by the opposing team during an attack by the observed team.

The proposed explanatory model aims to optimise the one described in a recent paper by Hernández and Anguera (2001). However, in contrast to these authors the analysis carried out in the present research has considered two teams facing each other, that is, interaction. In our view, this is important as the positioning of a team takes on greater value if the description of its players' spatial distribution is complemented by a description of the opposing team's positioning, thus enabling a more comprehensive and contextualised understanding of what happens during the game. In one way or another, all related units in the match are thus included in the description of the game.

On the basis of the results described, intervention in the game – specifically with respect to the team involved in the interaction – would prove to be highly advantageous. We are now at the point of being able to produce guidelines for reading, interpreting and putting into practice the motor behaviours which the players of one team should aim for, in unison, with respect to the interaction at a given point of the match. These guidelines inevitably imply teamwork, a collective involvement and commitment in the search for a shared objective: winning.

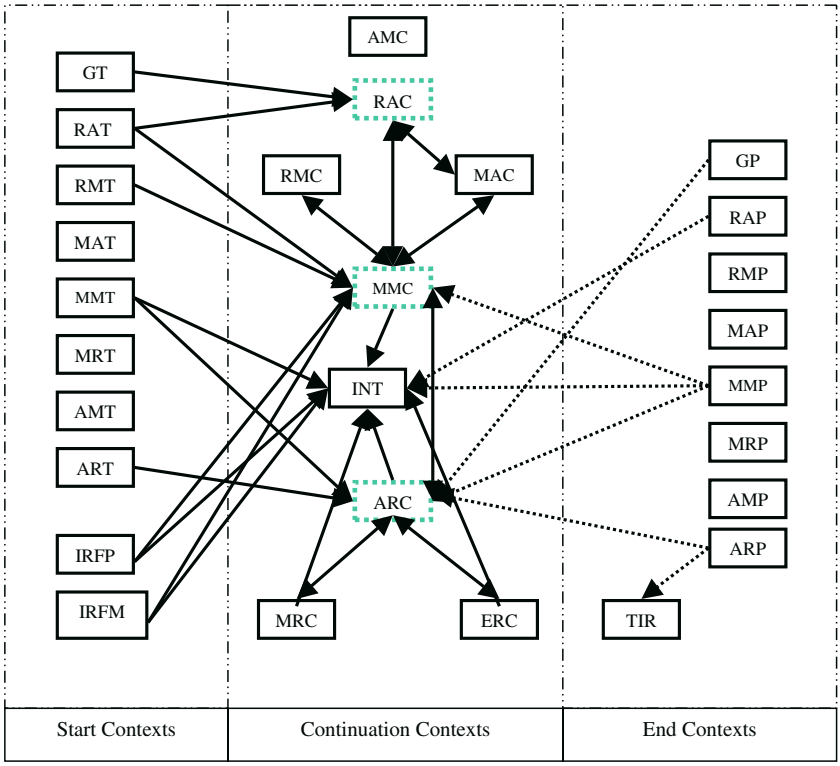

Figure 2. Proposed probabilistic model of play in a soccer match. In order to better understand the figure: (a) practically the same interaction contexts are repeated in the three situations into which play was divided: start, continuation and end of possession; (b) the contexts which have a discontinuous square have an implicit excitatory transition between them; and (c) the pointed arrows refer to the transitions from a retrospective point of view.

Naturally, the present model needs to be verified through further analyses. Its aim is to propose the sequences which characterise the system and generate a behavioural response of the teams with a certain degree of stability during a short period of time. The configuration of the model is marked by the spatial configurations of interaction generated and transformed by the two competing teams. These collective aspects must be collectively integrated into the model in order to provide a collective solution to the situations which occur continuously during the game.

References

Anguera, M. T. (1990). Metodología observacional. In J. Arnau, M. T. Anguera y J. Gómez Benito, *Metodología de la investigación en ciencias del comportamiento*. University of Murcia, Murcia, pp. 125–236.

Bakeman, R. & Quera, V. (1996). *Análisis de la interacción. Análisis secuencial con SDIS y GSEQ.* Madrid: RA-MA.

Castellano, J. & Hernández, M. A. (1999). Análisis secuencial en el fútbol de rendimiento. *Psicothema, 12* (supl. 2): 117–121.

Castellano, J., Hernández, M. A., Gómez de Segura, P., Fontetxa, E. & Bueno, I. (2000). Sistema de codificación y análisis de calidad del dato en el fútbol de rendimiento. *Psicothema* 12(4): 635–641.

Chappuis, R. & Thomas, R. (1988). *El equipo deportivo.* Paidós, Barcelona.

Hernández, M. A. & Anguera, M. T. (1997). Aportaciones del análisis secuencial a las acciones de juego en deportes sociomotores. *Proceedings of the V Congress on Methodology in the Human and Social Sciences*. Kronos, Seville, pp. 53–58.

Hernández, M. A. & Anguera, M. T. (2001). Estructura conductual en deportes sociomotores: fútbol. *Revista de Psicología Social*, 16(1): 71–93.

McGarry, T. & Franks, I. M. (1996). Development, applications and limitation of the stochastic Markov model in explaining championship squash performance. *Research Quarterly for Exercise and Sport* 67(4), 406–415.

Parlebas, P. (1987). *Activités physiques et éducation motrice.* 4, Dossier EPS, Paris.

Sackett, G. P. (1987). Analysis of Sequential Social Interaction Data: Some Issues, Recent Developments and a Causal Inference Model. In: J. D. Osofsky (ed.), *Handbook of Infant Development.* Wiley, New York, pp. 855–878.

Ysewijn, P. (1996). *GT: Software for Generalizability Studies*. Mimeografía.