DOI 10.2478/v10078-09-0002-9

Evaluation of Tennis Match Data - New Acquisition Model

by
Niksa Djurovic[1], Vinko Lozovina[2], Leo Pavicic[3]

The purpose of this research was to analyze and interpret the latent (factor) area of a tennis match. The entities in this research make 128 tennis matches played at the 2007 and 2008 Grand Slams hard court surfaces. The variables were created by use of the official statistics kept by the IBM Software - IBM DB2 Universal Database. The original variables were standardized to the number of sets in a match. A factor analysis under a component model was conducted. The number of factors retained, which was determined by the G-K criterion, explained 83.38 % of the total variance. Five significant factors substantiated the hypothesis established in this paper. The first factor named Match Successfulness is determined by the total number of break points; break points won and received points. The second factor named First Serve Significance is determined by the total number of first serves and winning points after the first serve. The third factor named Serve Speed is determined by the average speed of the serve and the fastest serve. The fourth factor named Net Play is determined by the total net approaches as well as the winning points after net approaches which are directly dependent on the total number of serves. The fifth factor named Play Errors is determined by unforced and double-fault errors. Winning matches differentiate from the lost matches by a smaller number of unforced and double-fault errors; considerably better results of the first serve, maximum serve speed and the number of aces scored, high score of total break points and break points won. The facts that do not differentiate winning matches from the lost ones are: first serves total, first serve throw-in, winning points after the first serve, number of net approaches and winning points after net approaches. The classification results show that with a system of 15 variables it is possible to recognize 96.0% of lost and 96.9% of winning matches. The achieved results indicate that the official match statistics with a modified system of 15 selected variables can properly interpret and predict match successfulness. This enables creators of a match observation system to valorize and enhance it with new indices.

Key words: directoblimin rotation, match successfulness, grand slam tournaments, tennis latent dimensions

Introduction

While watching a match through a TV broadcast, we are used to filling in the match analysis with various information on the screen. This relates to the so-called statistics comprising score changes, successfulness of particular actions and special indicators of player successfulness (O'Donoghue, 2001). We were interested in the relation of real values of a group of standard indicators to the forecasts of player successfulness and final match outcome (Pollard et al. 2006). For that purpose, we administered an analysis of standard statistics of world tournaments, in order to explain latent structure of indicators based on intercorrelations. Modern computerised match recording and analysis technologies fa-

[1] - *Faculty of Kinesiology, Split University*
[2] - *Faculty of Maritime Studies, Split University*
[3] - *Faculty of Kinesiology, Zagreb University*

Authors submited their contribution of the article to the editorial board.
Accepted for pinting in Journal of Human Kinetics vol. 21/2009 on May 2009.

cilitate detailed analyses by coaches and players; although they are a rich source of information they, however, require competent and sophisticated interpretation. Therefore, they are less valuable if used for play comparison and analysis (Magnus & Klaassen 1999). For such analyses it is necessary to determine the value and reliability of particular manifasting and latent indicators in relation to play successfulness. We made an analysis which enabled us to determine relative significance of particular indicators for the match outcome on a sample of worlds elite best players, as well as latent structure of the statistics kept on these matches. The variables observed cover several aspects of the play, and include: serve performance (preciseness and speed), winners, unforced faults, net play, errors, receives, winning points after the opponent's serve, breaks and other. In the structure of match play, these aspects will reflect in the latent structure obtained by factor/component analysis of the statistics of all matches covered. Likewise, it is important to determine the discriminant value of particular variables and latent dimensions for the forecast (prediction) of match outcome (Frings, 2006; Pollard at al. 2006; Scheibehenne & Broder 2007). The purpose of this research was to make a selection of variables using official statistics, create **new** ones and try to explain the match structure. By determining the structure of latent dimensions, we tried to explain what makes a tennis player successful. A hypothesis was established on the presumption that based on the variables in manifesting and latent areas of the match, it would be possible to identify the structures such as: *match successfulness, first serve significance, serve speed, net play* and *play errors*.

Methods

Sample of matches

The entities in this research make 128 tennis matches played at the 2007 and 2008 Grand Slams hard court surfaces. All matches were played on the concrete surface in order to standardise playing conditions and to neutralise effect of the surface relating to the specialists in playing on grass or clay surfaces. The original results, according to the official statistics of matches and players kept, were used with approval by the IBM Organisation.

Sample of variables

The results were collected during 128 official tennis matches. The variables describing the match were taken from the official statistics kept by the IBM DB2 application. On the basis of original variables, derived variables were selected and calculated. The matches were very different by duration and the number of sets played. This disproportion was settled by taking average results by set of the respective match i.e. by dividing the recorded - officially registered - data with the number of sets played in that match.

Description of variables

MWIN; 0 indicates a match lost, 1 indicates a match won

MGEMSUCC; MGAMEWIN/MGAMETOTAL

MMAXSERVE; maximum speed of the match serve

MAVERAG1SERV; average speed of the match first serve

Ma1srSERV; number of the player's first serves in the match divided by the number of sets played during the same match

Ma1srSERin; number of the first serves that hit the required opponent's field divided by the number of sets played during the same match

ManACES; number of serves untouched by the opponent

MaDOUBFO; number of double-fault errors at the serve divided by the number of sets played during the same match

MaUNFERR; number of unforced errors in the match divided by the number of sets played during the same match

MaPOI1SRV; number of points won after the first serve divided by the number of sets played during the same match

MaRECIVPNT; number of receive points won divided by the number of sets played during the same match

MaBRKPTOT; number of break points won divided by the number of sets played during the same match

MaBRKWIN; number of break points won divided by the number of sets played during the same match

MaNETGTOT; number of net approaches and play divided by the number of sets played during the same match

MaNETGWIN; number of winning points after net approaches divided by the number of sets played during the same match

Methods applied - Data analysis

The statistics, arithmetic means and standard deviations, minimum and maximum result, skewness and curtosis were calculated. A factor analysis under a component model was conducted, the number of significant factors was determined by the Guttmann-Kaiser criterion, the *Directoblimin* rotation was conducted, and the factors were presented by structure and pattern matrices. The factor scores and statistics of the groups of matches won and lost were calculated. A variance analysis was applied, and the structure of differences between groups was presented by the structure of discriminate function. The component model SPSS 13.0 was used for data analysis.

Results and discussion

The following conclusions can be derived from the results of a descriptive analysis of variables in this research:

- an average speed of a fully hit serve in men's professional tennis on the concrete surface is 207.75 km/h;
- an average first serve speed is 185.02 km/h;
- 29.96 first serves are averagely hit during one set, and 18.41 of them are successful;
- 2.35 aces are averagely won, and 1.01 double-fault errors and 9.54 unforced errors averagely made by set, with averagely 13.01 points after the first serve;
- averagely 2.54 break chances occur, and 1.10 of them are averagely successful;
- during a set, there are 8.67 net approaches, 5.60 of them successful (Table 1).

The factor analysis resulted in selecting five factors which interpreted 83.38% of the variance. Commonalities of all variables range from .68 to .97, which makes this system of measures stable and reliable for further analyses (Table 2).

The first factor, which makes up 28.22% of the common variance, is defined by variables MaBRKPWIN, MaBRKPTOT, MaRECIVPNT, MGEMSUCC and MWIN. This factor is named ***MATCH SUCCESSFULNESS***, which implies to a great number of total break points, break points won and received points what has a direct effect on match successfulness. The player with a greater number of total break points won (MaBRKPWIN) is basically the match winner. Logically, a greater number of break points won directly means a greater number of games won, and indirectly a greater number of sets, which leads to the winning match, and explains high coefficients on variables MGEMSUCC and MWIN (O'Donoghue, 2001). The player with a higher number of MaRECIVPNT can, just with this element, achieve the highest MaBRKPTOT, have a better chance for high MaBRKPWIN, and a chance to win the match. *The second factor,* which makes up 22.33% of the variance, is defined by variables Ma1stSERin, Ma1stSERV and MaPOI1SRV. This factor is named ***SERVE SIGNIFICANCE***, which in a game depends on the total number of serves, total number of first serves and winning points after the first serve. The first serve is definitely the most important technical element in tennis. It is the only technical and tactical element the performance of which does not depend on the player. By a brilliantly hit first serve, match

Table 1

Descriptive statistics: mean and standard deviation, MIN and MAX results, skewness and kurtosis

	Minimum	Maximum	Mean	Std. Dev.	Skewness	Kurtosis
MMAXSERVE	188	235	207.75	8.498	.194	.123
MAVERAG1SERV	159	209	185.02	8.938	-.022	-.092
Ma1stSERV	2.75	44.50	29.9647	5.48074	-.420	1.655
Ma1stSERin	8.67	28.80	18.4147	3.57570	.216	-.226
ManACES	.00	7.80	2.3501	1.42875	.739	.461
MaDOUBFO	.00	3.25	1.0086	.64857	.510	-.128
MaUNFERR	1.33	19.33	9.5435	3.32246	.282	-.082
MaPOI1SRV	5.00	21.00	13.0130	2.81461	-.103	-.027
MaRECIVPNT	3.33	20.33	11.2598	3.37640	-.149	-.384
MaBRKPTOT	.00	8.33	2.5362	1.50015	.608	.629
MaBRKPWIN	.00	3.00	1.1003	.72477	.349	-.550
MaNETGTOT	1.33	41.80	8.6656	4.78539	2.246	10.042
MaNETGWIN	.67	23.40	5.6009	2.99462	1.646	5.607

Table 2

Matrix of structure, pattern, communalities and variances in % by factor

Variables	Structure					Pattern					
	1	**2**	**3**	**4**	**5**	***1***	**2**	**3**	***4***	**5**	<u>***h***</u>
	Success	**1Service**	**Speed**	Net G	**Foults**	**Success**	**1Service**	**Speed**	**Net G**	**Foults**	
MWIN	**.83**	.04	.20	.11	-.40	**.77**	.03	.11	.09	-.25	0.78
MGEMSUCC	**.88**	.09	.27	.00	-.39	**.81**	.04	.18	-.02	-.25	0.86
MMAXSERVE	.02	.01	**.85**	-.12	.05	-.05	-.03	**.85**	-.08	.07	0.74
MAVERAG1SERV	-.03	-.15	**.88**	-.05	.03	-.09	-.19	**.90**	-.06	.07	0.83
Ma1stSERV	-.09	**.83**	-.02	-.33	.55	-.10	**.76**	-.01	-.06	.38	0.87
Ma1stSERin	-.01	**.94**	-.16	-.33	.20	-.09	**.93**	-.17	-.07	.01	0.93
ManACES	.23	.18	**.81**	.05	-.09	.12	.20	**.79**	.13	-.05	0.71
MaDOUBFO	-.07	.08	.01	-.04	**.81**	.09	-.07	.03	.06	**.85**	0.68
MaUNFERR	-.21	.32	.04	-.21	**.79**	-.10	.18	.07	-.06	**.74**	0.68
MaPOI1SRV	.30	**.92**	.19	-.29	-.01	.17	**.91**	.14	-.04	-.14	0.93
MaRECIVPNT	**.88**	.30	.04	-.11	.06	**.91**	.16	-.05	-.05	.19	0.86
MaBRKPTOT	**.88**	.04	-.01	-.03	-.03	**.92**	-.09	-.10	-.05	.15	0.80
MaBRKPWIN	**.94**	.00	.05	.04	-.18	**.95**	-.09	-.05	.00	.01	0.89
MaNETGTOT	-.06	.28	.04	**-.99**	.13	-.06	.01	.01	**-.98**	-.02	0.97
MaNETGWIN	.12	.31	.07	**-.98**	.07	.11	.03	.02	**-.98**	-.05	0.97
% of Variance	***28.22***	***22.33***	***14.82***	***10.01***	***8.00***						

initiative is achieved, and the opponent is also psychologically defeated (Crognier & Féry 2005). An average first serve speed (Table 1) in men's senior category is around 207.75 km/h (Pugh et al. 2003). *The third factor,* which makes 14.82 % of the variance, is defined by variables MAVERAG1SERV, MMAXSERVE and ManACES. This factor is named ***SERVE SPEED.*** The total number of first serves directly depends on total match duration. Successfulness of the game/set/match is also directly dependent on the winning points after the first serve, while successfulness of the first serve depends on the high average serve speed, as well as high fastest serve (Pugh et al. 2003). A serve hit at a speed range of 208-235 km/h has a very high probabilty of being successful.

In cases where the serve speed exceeds 208 km/h, there is a great probability for scoring an ace (ManACES .81), or a directly won point without playing. *The fourth factor*, which makes up 10.01% of the variance, is mainly defined by variables MaNETGTOT and MaNETGWIN. It is further defined by significant negative projections of high values and associated variables of slightly lower, but still significant values of the same direction Ma1stSERin and Ma1stSEV. This factor is named ***NET PLAY***. The total net approaches, as well as the winning points after net approaches are directly dependent on the total number of first and second serves. Also, the total number of first serves and the number of throw-ins directly determines the number of net approaches and indirectly the winning points after net approaches. This type of play is preferred by serve-and-volley players; the strategy is to force the opponent to an error and try to return a hard passing shot (Chow et al. 1999). It is more efficient on faster surfaces due to faster bounce of the ball, and where the speed of play limits the opponent's time while preparing for a passing shot. *The fifth factor,* which makes 8% of the variance in the whole system, is defined by significant positive projections of variables MaDOUBFO and MaUNFERR, as well as associated variables of slightly lower, but still significant opposite direction values MWIN and MGEMSUCC (-.39), and by a positive direction variable Ma1stSEV. This factor is named ***PLAY ERRORS***, and is defined by unforced and double-fault errors. Unforced errors usually result from a wrong decision or poor hitting technique (Brody 2006).

Factor intercorrelations are low. The only significant, but negative correlation is between the second and forth factors (r = -.28), which can be explained in the following way: the total of net approaches and the number of winning points after net approaches directly depends on the number of first and second serves. Low intercorrelations of other factors support the fact that each of them exists independently. Factor scores were calculated for lost and winnning matches. The score of differences was calculated on

Table 3

Statistics by factor scores, variance analysis, equation test of group arithmetic means, standardised coefficients of the canonical discriminant function and its structure

Factor	Match lost		Match win		Wilks' Lambda	F	Sig.	Disc. Function Struct.	Coefficients
	Mean	Std Dev.	Mean	Std Dev.					
SUCCESS	**-.840**	.583	**.827**	.519	**.303**	**579.448**	**.000**	**.812**	1.024
1SERVICE	-.040	1.066	.039	.933	.998	.397	.529	.021	.074
SERV. SPEED	**-.200**	1.000	**.197**	.965	**.961**	**10.334**	**.001**	**.108**	.263
NETGAME	-.114	1.138	.113	.832	.987	3.298	.071	.061	.205
ERORS	**.401**	.915	**-.395**	.923	**.841**	**47.639**	**.000**	**-.233**	-.540

N=253, Df1=1, df2=252

the first discriminant function where the Wilks Lamda was 0.22, and the tested value of Chi-square was 374.6 which, at 5 degrees of independence, is statistically significant (p≤0.001). The first, third and fifth factors have significant discriminant power of differentiating the lost from winning matches. The second and fourth factors, however, do not differentiate the winning from lost matches (Table 3). The winning matches are characterised, and differ from the lost ones, by a lower number of double-fault and unforced errors on the **fifth factor**, as well as significantly better results of the average first serve speed, maximum serve speed and the number of aces scored on the **third factor**. The **first factor** is characterised by a high score of break points and winning break points, the total of receives regardless of whether the point was scored after the first return or after several interchanges, and the number of games won, relatively to the total, as well as to the winners total. The facts that do not differentiate the winning from lost matches are the following: in the **second factor** - the first throw-in, points won after the first serve, the total of first serves hit, and in the **fourth factor** - the number of net approaches, as well as the points won after net approaches. The classification results indicate that, with a system of 15 variables retained for final analysis, it is possible to recognise 96.0% of lost and 96.9% of winning matches, which practically means that the system of retained variables functions perfectly in match analysis (Tables 2 and 3) (Newton & Keller 2005).

Conclusion

In compliance with the purpose of research, 15 variables were selected from the official IBM statistics. In order to standardise the matches, the variables were divided by the total number of sets played and one of them was divided by the games total. The statistics of all variables were calculated, and a factor analysis under a component model was conducted. The number of factors was determined by the G-K criterion, and five obtained factors interpreted 83.38% of the system variances. In the latent area, the first factor, ***successfulness***, is defined by the number of break points total and win after the opponent's serve. The second factor, ***first serve significance***, explains that the game success depends on the total number of first serves and total winning points after throw-ins. The third factor, ***serve speed***, explains that successfulness of the first serve depends on the high average serve speed, as well as on high fastest serve. The fourth factor, ***net play***, explains the total net approaches as well as winning points after net approaches. The fifth factor, ***play errors***, is defined by unforced and double-fault errors. Factor scores were calculated, a variance analysis was conducted, and the discriminant function structure was calculated. Winning matches are characterized, and differentiated from the lost matches, by a smaller number of double-fault and unforced and errors, considerably better results of the average first serve speed, maximum serve speed and number of aces scored, as well as a high score of break points total and won, total received points no matter if scored after first returns or after several interchanges, and the number of games won relative to the total. The facts that do not differentiate winning matches from the lost ones are: i**n the second factor** - the first serve throw-in, winning points after the first serve and total first serves hit; i**n the fourth factor** it is the number of net approaches and winning points after net approaches. With a modified system of 15 variables based on the official IBM statistics, it is possible to recognize 96.0% of lost and 96.9% of winning matches, which practically means that a system of five latent dimensions perfectly functions for the purpose of match analysis. The results presented in this paper will facilitate future creators of a match observation system to conduct valorization of the existing system and possibly enhance it with new indices.

References

Brody H. Unforced errors and error reduction in tennis. Brit J Sports Med, 2006. 40: 397-400.

Bruce E. Biomechanics and tennis, Brit J Sports Med, 2006. 40: 392-396.

Chow J.W., Carlton L.G., Chae W., Shim J., Lim J.,Kuenster, A.F. Movement characteristics of the tennis volley. Med Sci Sports Exerc, 1999. 31: 855-863.

Crognier L., Féry Y.A. Effect of Tactical Initiative on Predicting Passing Shots in Tennis. Appl Cognitive Psych, 2005. 19: 637- 649.

FischerG. Exercise in probability and statistics, or the probability of winning at tennis. Am J Phys, 1980. 48(1), 14-19.

Frings C. Who will win Wimbledon? The recognition heuristic in predicting sports events. J Behav Decis Making, 2006. 19: 321-332.

IBM Software - IBM DB2 Universal Database.

Klaassen F., Magnus J. Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. J Am Stat Assoc, 2001. 96: 500-509.

Magnus J., Klaassen F. The final set in a tennis match: Four years at Wimbledon. J Appl Stat, 1999. 26: 461-468.

Magnus J., Klaassen F. On the advantage of serving first in a tennis set: Four years at Wimbledon. The Statistician, 1999. 48: 247-256.

Match Analysis DVD, http://www-03.ibm.com/press/us/en/pressrelease/24991.wss

Miles R.E. Symmetric sequential analysis: the efficiencies of sports scoring systems. J Roy Stat Soc B Met, 1984. 46: 93-108.

Newton P.K., Keller J.B. Probability of winning at tennis I. Theory and data. Stud Appl Math, 2005. 114: 241-269.

O'DonoghueP.G. The most important points in Grand Slam singles tennis. Res Q Exercise Sport, 2001. 72: 125-131.

Paserman M.D. Gender Differences in Performance in Competitive Environments: Evidence from Professional Tennis Players, 2007. C.E.P.R. 6335 Discussion Papers.

Pollard G., Cross R. , Meyer D. An analysis of ten years of the four grand slam men's singles data for lack of independence of set outcomes. J Sports Sci Med, 2006. 5 : 561-566.

Pugh S. F., Kovaleski J. E., Heitman R. J., Gilley W. F. Upper and lower body strength in relation to ball speed during a serve by male collegiate tennis players. Perceptual and motor skills, 2003. 97 : 867-872 .

Riddle L. H. Probability Models for Tennis Scoring Systems . Appl Stat, 1988. 37: 63-75.

ScheibehenneB., Broder A. Predicting Wimbledon 2005 tennis results by mere player name recognition. Int J Forecasting, 2007. 23: 415-426.

Acknowledgments

We thank IBM for their kindness in providing the data. We are also grateful to Ljubica Stanisic and Jadranka Majic for their careful help in translating this article to English version.

Corresponding author

Prof. **Niksa Djurovic**

University of Split, Faculty of Kinesiology
Simiceva 9, 21000 Split, Croatia
Phone/fax: +385 98 164 2903
E-mail: niksa.djurovic@gmail.com
niksa@kifst.hr