

Franc J.G.M. Klaassen, PhD (1) and Prof. Jan R. Magnus, PhD (2)

1. Department of Economics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. Tel: +31-20-5254191, Fax: +31-20-5254254. E-mail: f.klaassen@uva.nl.

2. CentER for Economic Research, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands. Tel: +31-13-4663092, Fax: +31-13-4663066. E-mail: magnus@uvt.nl.



Franc Klaassen

During a tennis match broadcast on TV, a number of interesting statistics are presented to the viewers. The most obvious one is the score, but the percentage of first serves in, the number of aces, and a few other statistics are also regularly reported on TV. These statistics are then discussed by the commentators to provide a deeper insight into various aspects of the match. However, a direct statistic concerning the most important aspect of the match- namely who will win - is not shown. In this article we present a method to compute such a statistic: i.e. the probability that a given player will win the match.

There are already several methods for calculating that probability at the start of the match. One could use the odds from bookmakers. Alternatively, one could employ a statistical model, such as the model developed by Clarke and Dyte (1) who use official (ATP and WTA) rating points to estimate the probability that a player will win. In a match between players A and B, for instance, this could be 70% for player A. However, as the match progresses, new data become available, and these can be used to update the pre-match probability. For instance, if A has lost the first set, then the probability that A wins drops: the

question is by how much. This article describes the method of Klaassen and Magnus (1) to estimate this probability. More specifically, we show how one can compute the probability of A winning, not only at the start of the match, but also (and in particular) at each point during the match. This results in a graph of subsequent winning probabilities, which unfolds during the match. If the estimate exceeds 50% for a player, then that player is predicted to win the match. Hence, the graph also gives forecasts for the winner of the match.

The graph and underlying probabilities can be informative for TV viewers watching a tennis match. After all, the score, although implying who currently leads the match, does not give a perfect indication of the likely winner of the match: a top player may still be the favorite after losing the first set. The score also gives only partial information on the development of the match: a score of 5-5 can result after 4-4, but also after 5-0. Summary statistics, such as the percentage first serves in and the number of aces, do not contribute much in these respects (which is not surprising, because their main objective is to provide insights into the way players play the points). An estimate of the probability that A wins the match, however, provides a direct indication of the likely winner of the match. In addition, the graph of probabilities at all points played so far gives an overview of the match development up to now; in fact, it makes the information visible at a glance, so that it may be useful to project the graph on TV and let it support the commentator in his/her discussion of the match.

Method

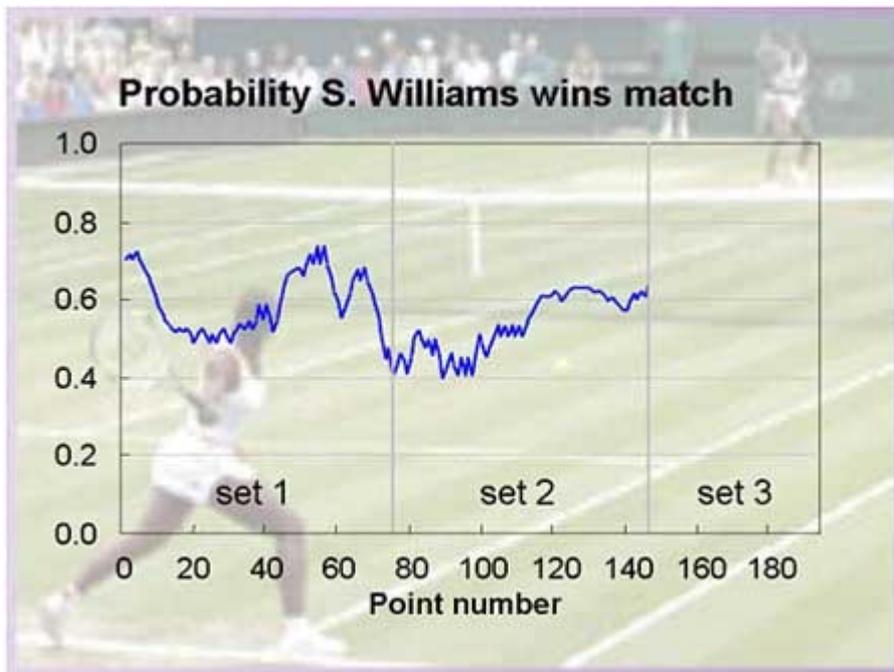
To discuss the computation of the probabilities and thereby the complete graph, we distinguish between the pre-match probability (the first point of the graph), and the updated probabilities during the match (rest of the graph). To estimate the first probability, we suggest using a transformation of the official rankings. This leads to a probability of (e.g.) 80% that player A will win against B. Of course, rankings are just one indicator of the relative strength of the players. If other information is available, on special abilities on the court surface and injury problems, for example, then adjustments can be made and possibly the ranking-based estimate can be improved. Ultimately, there will be one pre-match estimate of e.g. 70%. Klaassen and Magnus (3) show that taking another reasonable pre-match estimate shifts the graph somewhat, but leaves its form practically unchanged. Hence, the usefulness of the graph in practice does not depend on the exact starting point.

To update the estimate as the match progresses, we have written a computer programme called 'Tennisprob'. Given the rules of the tournament (best-of-three-sets or best-of-five-sets match, tiebreak in final set or not), the current score, the current server, given the statistical assumption that winning points on serve is an independent and identically distributed process (see Klaassen and Magnus2 for a justification of this assumption), and given two input probabilities, 'Tennisprob' computes the probability that A wins the match at the beginning of the point under consideration. This probability is computed exactly (not by simulation) and very fast (within a second). The first input probability is the pre-match probability estimated above. The second input consists of the sum of two

point probabilities: the probability that A wins a point on serve and the probability that B wins a point on serve. As before, we have developed a method that uses the rankings to estimate this sum. This leads to e.g. 120%. This estimate needs no further adjustment, because the probability of interest (that A wins the match) hardly depends on it. These two pieces of information are all we need. In particular, we do not need to estimate the two point probabilities - just their sum and the initial match probability are sufficient. This is important, because the latter probabilities are more robust to misspecification. Note also that no information on the future development of the match is needed.

To show how the forecasting procedure works in practice, let us consider the 2003 women's singles Wimbledon final between Serena and Venus Williams. Before the match starts, we have to choose the two input probabilities introduced above. The first one is the probability that Serena (player A) wins the match before any point has been played. Because Serena's WTA ranking was 1 and Venus' ranking was 4, we obtain a pre-match estimate of 81%. However, the knowledge that Venus had reached the Wimbledon final for the last three years and had won two of them means that the pure ranking based estimate was probably too high. On the other hand, Venus was injured. Taking this information into account, we believed that a reasonable starting probability is 70%. Of course, one may put more weight on the injury aspect and prefer a higher probability; in that case the starting point has to be adjusted upwards. The second input probability for 'Tennisprob' is the sum of both players' probability of winning a point on service. Our ranking-based estimate is 116%, implying an average 58% probability of winning a point on service. This seems reasonable.

With these two inputs the probability that Serena will win the match can be computed at each point in the match. In fact, once a point is finished and the new score is known, the updated probability is presented within one second. Suppose the second set of the final has just been completed (the score from Serena's point of view is 4-6/6-4). Then the following graph can be generated.



The graph shows that at the start of the final set (145 points have been played) Serena's probability has shrunk from the starting point of 70% to about 60%. The graph also gives a summary of the development of the match so far. Venus' start was overwhelming. Serena came back but still lost the first set. The first few games of the second set were shared, but then Serena took a decisive lead. In the second part of the second set the graph is quite stable, indicating that halfway through the second set, it was already quite certain that Serena would win the set. It took Serena some time before she could end the second set. During the third set the graph can be extended.

Conclusion

The Serena-Venus match is just one example. A graph can be produced for any match for which we have the two pre-match input probabilities and the point-to-point information as the match unfolds. Therefore, the approach described above is a generally applicable forecasting method. By giving information on the likely winner and the past development of the match, the graph gives extra information besides the score and the summary statistics that are commonly presented on TV. It also makes this information visible at a glance and the graph can be generated instantly. Hence, it may be interesting to show the graph on TV at the changes of

ends. Commentators could then use the graph to discuss the match, and also to evaluate the match afterwards.

References

1. Clarke SR, Dyte D. Using official rating to simulate major tennis tournaments. Intern Trans Oper Res 2000;7:585-594.
2. Klaassen FJGM, JR Magnus. Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. J Am Stat Assoc 2001;96: 500-509.
3. Klaassen FJGM, JR Magnus. Forecasting the winner of a tennis match. Eur J Oper Res 2003; 148: 257-267.

About the authors

Franc Klaassen is Associate Professor of Economics at the University of Amsterdam in The Netherlands. He obtained his Ph.D. at Tilburg University in the Netherlands in 2000. His main research interests are empirical international economics and finance (for instance, exchange rates, exports, effects of the Euro on trade in the European Union), econometrics and the statistical analysis of sports, mainly tennis.

Jan Magnus is Research Professor of Econometrics at Tilburg University in The Netherlands. He obtained his Ph.D. at the University of Amsterdam in the Netherlands in 1981. His research interests include econometrics, both methodology and applications, matrix calculus, statistics, national accounts estimation and sport statistics, in particular tennis.

