DATA DOCUMENTATION FOR: Disequilibrium Play in Tennis

Axel Anderson, *Georgetown University*[*]
Jeremy Rosen, *Georgetown University*[†]
John Rust, *Georgetown University*[‡]
Kin-Ping Wong, *Digonex*[§]

March 9, 2024

[*]Department of Economics, Georgetown University, e-mail: aza@georgetown.edu
[†]Department of Economics, Georgetown University, e-mail: jar361@georgetown.edu
[‡]Department of Economics, Georgetown University, e-mail: jr1393@georgetown.edu
[§]Digonex, e-mail: kjw65@georgetown.edu

1 Overview

We obtained the data from the Match Charting Project (MCP) http://www.tennisabstract.com/charting in 2018 (with permission from the site). Our data include 3,587 tennis matches played in 2,663 tournaments; the first match was in New South Wales on March 22, 1970, and the most recent was in Shenzhen on January 5, 2018. The data we are providing complies with the Creative Commons License for the Match Charting Project that is available at:

`https://github.com/JeffSackmann/tennis_MatchChartingProject?tab=readme-ov-file`
`https://creativecommons.org/licenses/by-nc-sa/4.0/`

The Appendix includes an email reply from the founder, Jeff Sackman, confirming that the posted data are in compliance, and there is no issue in redistributing the small subset of the MCP we downloaded in 2018 from tennisabstract.com to be parsed and re-posted on the JPE Dataverse site for purposes of data replication of the results in "Disequilibrium Play in Tennis," Journal of Political Economy, 2024 (forthcoming).

The MCP uses volunteers who watch each point of each match and record shot by shot descriptions of the points. In each "Point-by-point description" page, the server is shown in the first column, the set score is in the second column, the game score in the current set is in the third column, and the point score in the current game is in the fourth column. The fifth and final column is a written summary of the type and outcome of each shot. For all the 3,587 matches we obtained, there are a total of 958 distinct players and 548,302 point by point descriptions. Below is an example of Andre Agassi serving a game to Pete Sampras in the Jan 29, 1995 Australian Open final, where we show the fourth and fifth columns.

```
0-0    | 1st serve wide; forehand return down the middle (shallow);
         forehand crosscourt; forehand crosscourt; forehand crosscourt;
         forehand crosscourt; forehand crosscourt (net cord); forehand crosscourt;
         forehand down the line (wide),  unforced error. (8-shot rally)
0-15   | 1st serve wide,  fault (net). 2nd serve wide; backhand slice
         return crosscourt (long), unforced error.
15-15  | 1st serve wide,  fault (net). 2nd serve to body; forehand return down
```

```
          the middle (shallow); forehand crosscourt; forehand down the line;
          backhand crosscourt; backhand crosscourt; forehand inside-out;
          backhand crosscourt; forehand inside-in; forehand crosscourt (net),
          unforced error. (9-shot rally)
 30-15  | 1st serve wide,  fault (wide and long). 2nd serve wide; backhand return
          crosscourt (deep); forehand inside-in,  winner.
 40-15  | 1st serve down the T; backhand return crosscourt (deep); backhand crosscourt;
          backhand down the middle (long), unforced error.
```

We parsed and recoded this data into a purely numerical format. The data are stored in a plain text file, named `lcrdata.txt`, and we also appended as an eighth column the type of surface in each game, with: 1 for hard court, 2 for clay court, and 3 for grass court.

The states of tennis can be regarded as states of a Markov chain with possible transitions as diagrammed in the tree. There are two absorbing states in this chain: State 37 (win for the server) and State 38 (loss for the server). State 1 is the starting state with a score of 0–0. State 2 denotes a second serve at 0–0 in the event of a faulted first serve. Our encoding is set up so that the state number for all first serve states is odd and for all second serve states is even, which is a helpful way to determine if the current state is a first or second serve. Our numerical codes for the tennis states are shown in figure 1 of the paper, "Disequilibrium Play in Tennis" and are also stored in the table `game_state_definitions` in the tennis database.

Below are the first eight lines of `lcrdata.txt`, which contains our numerical encodings for each match in the data we used to estimate the model for 46 elite professional server-receiver pairs on hard surfaces that we analyzed in sections 4 and 5 of the paper. We also provide an additional file `lcrdata_extra.txt` that contains data on 114 additional player pairs who played matches on hard, clay and grass surfaces that were used for our analysis of additional player pairs in section 5.3 of the paper.[1]

[1]Note that not that we did not get a valid variance-covariance matrix for the POPs for all 114 of these pairs (and we need this covariance matrix to generate the random perturbations of the POPs for our robust analysis of deviation gains in section 5.3), so section 5.3 presents the results for a subset of 99 of the 114 cases where the variance-covariance matrix is valid. Due to convergence and sample size issues, the maximum likelihood estimator would not always result in a non-singular information matrix (or hessian matrix), and hence we were unable to invert it in these cases to obtain an estimate of the variance-covariance matrix of the parameters of our model of the POPs, which we used to generate random perturbations of the POPS about the maximum likelihood point estimate of the POPs using this covariance matrix.

```
1,1,3,2,9,0,168,1
1,9,1,3,10,0,168,1
1,10,1,1,11,7,168,1
1,11,3,3,12,7,168,1
1,12,2,1,13,3,168,1
1,13,1,3,14,3,168,1
1,14,1,1,15,7,168,1
1,15,1,1,37,7,168,1
```

The first column is the server-returner ID, which is 1 for Agassi serving to Sampras. The ID code of the server-returner pair can be found in the file `all_players_list.txt` and also the field `pair_id` of the table `player_pairs` in the tennis database. The `lcrdata.txt` file was produced from all matches on hard courts for player pair IDs 1 to 46. In addition, our analysis in section 5.3 of the paper was based on an additional data file `lcr_data_extra.txt` that was created from player pair IDs 1001 to 1114 in the `player_pairs` table.

Note that since service alternates across games in a set, the `lcrdata.txt` file only picks up the games in the sets of various matches where Agassi was the server and Sampras the receiver. There is a `game_id` code in the `lcrdata` table of the tennis base that is an incremented counter of each service game played in any given tennis match. Thus, for a given server/receiver pair we only use either the even or odd game IDs, respectively, depending on whether the outcome of a coin flip at the start of the match decides whether the designated server in each given server/returner pair won the flip and was the server for the first game in the match, or for the second game (and thus continued to serve in the odd or even game IDs in that match).

The second column is the state of the game from 1 to 38. The third column is the serve direction, which is encoded as 1 for a serve to the *receiver's* left, 2 for a server to the receiver's body, and 3 for a serve to the receiver's right. NOTE: Since tennis serves alternate between the deuce and ad courts, the a wide (T) serve is to the receiver's right (left) when the server serves to the deuce (ad) court. For example, in the first serve in the above example, Agassi served wide, which is to the receiver's right, hence the serve direction is coded as 3. Meanwhile, the first serve at State 9 (0–15) is to the ad court, so this wide serve now goes to the receiver's left, meaning the serve direction is coded as 1.

The fourth column is the serve outcome, which is coded by three integer values:

1 denotes a successful serve where the server wins the point,

2 denotes a successful serve where the server loses the point, and

3 denotes a faulted serve.

In any first serve state, a fault results in a transition to the associated second serve state. For example, in the first serve at State 9 (0–15), Agassi faulted, so the outcome is coded as a 3, and the game transitions to State 10 (0–15, second serve). The fifth column is the state the game transitions to. In any second serve state, a fault results in the loss of the point, i.e. a double-fault.

The sixth column is a "muscle memory state" m that takes on nine possible values. It encodes the directions chosen by the server in the *two previous first serves.* Call the serve direction two first serves ago sd2, and the serve direction on the previous first serve sd1, and let each of these directions take the integer values 1 to 3 according to the direction encoding discussed above. Then m=3(sd2-1)+sd1. The reason we want to track the two previous first serves is due to the alternation of serves between the deuce and ad courts. We hypothesize that a server knows whether the receiver is right or left handed and whether the receiver is relatively stronger or weaker in returning serves hit to different directions from the receiver's perspective.

We assume that muscle memory is initialized to null (0) at the start of each game, so the muscle memory actually takes on 10 possible values once we account for this null value at the start of each game. However, in the example below, by the second serve at State 10 (second serve for Agassi at score 0–15), we have m=7=3*(sd2-1)+sd1=3*(3-1)+1. We do not update the muscle memory state after any second serves, and so the muscle memory at State 11 (15–15) is also 7.

The seventh column is an integer ID for each match, which traces back to the particular url of the MCP database where we obtained the data. In this case, 168 maps to the url:

`https://www.tennisabstract.com/charting/`
`19950129-M-Australian_Open-F-Pete_Sampras-Andre_Agassi.html`

This match index is stored in the field ID in the table `masterlist` of the Postgres tennis database, a compressed dumpfile which is also uploaded to the Dataverse site. Besides the masterlist table, another key table is players, which has the names and ID codes for the 958 distinct

professional tennis players in the data we downloaded from www.tennisabstract.com. In addition, the table point_by_point_description contains a copy of the point by point descriptions of each game in each set of each match downloaded from www.tennisabtract.com. There is a integer field called point_seq that orders the data according to the ordering of the information in the point by point descriptions we downloaded from www.tennisabstract.com. For example, the postgres query:

```
select * from point_by_point_description where game_id=168 order by point_seq;
```

displays all serves of every game of every set in the match between Pete Sampras and Andre Agassi in the 1995 Australian Open in the same order as the www.tennisabstract.com site at the url given above. NOTE: It might be possible that there have been updates or corrections on the www.tennisabstract.com website since we downloaded this data in 2018. Thus, we cannot guarantee there is an 100% match between the data we downloaded then and what is on the www.tennisabstract.com site currently. However, we provide the dump of the data we downloaded in 2018 for replication purposes, and the coding of the verbal point by point descriptions in the tennis database dump on the Dataverse site does result in the numerical encoding of serve directions and point by point outcomes for the subset of matches we analyzed, as they are in the lcrdata.txt file.

The eighth and final column of lcrdata.txt is an integer index for the court surface:

1 for hard court,

2 for clay court,

3 for grass court.

Most of our observations are from matches on hard courts: of the 5,951 games played by the 46 elite professional players our dataset, 3,516 or nearly 60% were played on hard courts.

Summary of the 8 columns in lcrdata.txt

1 server-returner ID (1 to 46, for elite pro player pairs on hard surfaces, 1001 to 1114 for other professional players on hard, clay, and grass surfaces included in the file *lcrdata_extra.txt*)

2 game state (1 to 38), see figure 2 of the paper for definitions, or table `game_state_definitions` in the tennis database.

3 serve direction (1=left, 2=body, 3=right) (from receiver's perspective)

4 serve outcome (1=serve in, server wins point, 2=serve in, receiver wins, 3=faulted serve)

5 new state to which game transits after the point outcome

6 muscle memory states (0 for initial state, or 1 to 9 encoding the two previous *first serve* directions, $(sd2, sd1)$ as $m = 3(sd2-1)+sd1$ except for the first serve of each game where m is initialized to 1, and in each 2nd serve of each game where $m = sd1$. Note that muscle memory is not updated after any second serve in a game, so $sd2$ and $sd1$ only refer to serve directions on *first serves* in the game.

7 match ID: a unique identifier of each match in our database. For further information on the match including the URL to the original data on `tennisabstract.com` that we downloaded the match data from, the date and location of the match, the names of the players and other information, see the table `masterlist` in the tennis database.

8 court surface: 1: hard, 2: clay, and 3: grass

Note that `lcrdata_extra.txt` follows this same format except the 8th column contains the year of the match, not a code for the court surface. However the surface code (and other match-specific information) can easily be obtained for the matches included in the `lcrdata_extra.txt` table by looking up the surface code in the `masterlist` table in the tennis database for each match ID.

2 Programs used to create the database

3 Appendix: Summary of Data Files

- `data_explanation.pdf` this file, providing the overall explanation of the uploaded data and program files

- READ.ME a plain text file directing the user to view this file for an overview of the data and programs that are uploaded to this site.

- tennis.dump a compressed Postgres dump file of the full database collected from the Match Charting Project for this study, with primary text data on outcomes of tennis matches in the point_by_point_description table. First stage processing of these textual descriptions resulted in numerically encoded outcomes of each serve in this table that was stored in the master_pointlog table. Second stage processing then further encoded into the lcrdata table that contains the data used in the empirical analysis. The tennis database also includes additional tables: 1) game_state_defintions that define our integer encoding of the 38 game states of a service game in tennis (i.e. point scores including win (37) and loss (38) by the server), 2) players a list of the names and ID codes for all tennis players included as a in any match in the data we obtained from tennisabstract.com, 3) masterlist a table providing metadata on the 3587 tennis matches that we downloaded from tennisabstract.com including the url to the data on that site, the names and ID codes of the players in each match, dates and location of the match, the court surface, and other information, 4) player_pair a list of the 46+114=160 distinct player pairs we analyzed in the paper, where pair ID codes from 1 to 46 index elite pro player pairs playing on hard surfaces that were used to generate the estimation data file lcrdata.txt and pair ID code 1001 to 1114 index 114 additional player pairs used to generate the estimation data file lcrdata_extra.txt.

- lcrdata.txt a plain text file containing data extracted from the lcrdata table of the tennis database for player pairs 1 to 46 in the player_pairs table of the tennis database, indexing elite pro players playing on hard surfaces that was used for the main part of the empirical analysis in the paper in sections 4 and 5.

- lcrdata_extra.txt Additional data used to estimate the model for an additional 99 player pairs analyzed in section 5.3 that include non-elite women and men pros playing on grass, clay and hard surfaces. Note that this data file actually contains data on 114 player pairs with player pair IDs 1001 to 1114 that were extracted from the lcrdata table of the tennis database, but estimations for all 114 of these pairs produced non-singular information (or

negative loglikelihood hessian) in 99 cases. As described in footnote 1 above, we needed to restrict to player pairs where the information matrix for the estimated POP parameters is non-singular in order to simulate random POPs from the asymptotic distribution of POPs, centered at the maximum likelihood estimate of each player pair's POPs.

- `all_players_list.txt` a list of the names and pair ID codes for the 46 elite tennis pros playing on hard surfaces that were analyzed in sections 4 and 5 of the paper. This is an extract from the table `player_pairs` in the `tennis` database.

- `all_players_list_extra.txt` a list of the names and pair ID codes for the 99 elite tennis pros playing on hard surfaces that were analyzed in sections 4 and 5 of the paper. This is an extract from the table `player_pairs` in the `tennis` database for player pair IDs 1001 to 1114. As discussed above though there are 114 player pairs in this table and the empirical analysis in section 5.3 reports results for 99 of these 1114, this discrepancy is due to the inability to obtain non-singular information matrices for all 114 player pairs, which we need to construct simulated draws from the asymptotic distribution of the POPs.

4 Appendix: Compliance with Creative Commons License

Below is an email reply from Jeff Sackman, developer of the `tennisabstract.com` website where we obtained our data, dated February 26, 2024 confirming that the MCP data we are posting here comply with the `tennisabstract.com Creative Commons License at:`

`https://github.com/JeffSackmann/tennis_MatchChartingProject?tab=readme-ov-file`
`https://creativecommons.org/licenses/by-nc-sa/4.0/`

`---------- Forwarded message ---------`
`From: Jeff Sackmann <jeffsackmann@gmail.com>`
`Date: Mon, Feb 26, 2024 at 1:19AM`
`Subject: Re: Use of Match Charting Project data for academic research`
`To: Jeremy Rosen <jar361\@georgetown.edu>`

Thanks Jeremy, that’s all fine. I’ll take a look at the paper.

On Mon, Feb 26, 2024 at 5:15AM Jeremy Rosen <jar361\@georgetown.edu> wrote:

Dear Mr. Sackmann,

My coauthors (Axel Anderson, John Rust, Kin-ping Wong) and I used Match Charting Project data for our academic paper "Disequilibrium Play in Tennis," which was accepted for publication at the Journal of Political Economy.

The journal requests that the data we used be posted on their website as part of a replication package; a link to the page is here:

https://doi.org/10.7910/DVN/RQ6JVL

We believe we’ve complied with the Match Charting Project’s Creative Commons license, but we just want to make sure that everything we’re doing is okay with you.

Also, here’s a link to our paper if you’re interested:

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4383716

as well as a pdf copy that contains on Page 69 a section called "Data Availability" that summarizes the data we provide for replication, which we had originally obtained from the Match Charting Project.

Sincerely,

Jeremy