

TEXTBOOK

Daniel Memmert *Editor*

Computer Science in Sport

Modeling, Simulation,
Data Analysis and Visualization
of Sports-Related
Data

FLASH-
CARDS
INSIDE

MOREMEDIA



Springer

Computer Science in Sport

SPRINGER NATURE



SN Flashcards Microlearning

Quick and efficient studying with digital flashcards – for work or school!

With SN Flashcards you can:

- **Learn** anytime and anywhere on your smartphone, tablet or computer
- **Master** the content of the book and test your knowledge
- **Get motivated** by using various question types enriched with multimedia components and choosing from three learning algorithms (long-term-memory mode, short-term-memory mode or exam mode)
- **Create** your own question sets to personalise your learning experience

How to access your SN Flashcards content:

1. Go to the **1st page of the 1st chapter** of this book and follow the instructions in the box to sign up for an SN Flashcards account and to access the flashcards content for this book.
2. Download the SN Flashcards mobile app from the Apple App Store or Google Play Store, open the app and follow the instructions in the app.
3. Within the mobile app or web app, select the flashcards content for this book and start learning!

If you have difficulties accessing the SN Flashcards content, please write an email to customerservice@springernature.com mentioning “SN Flashcards” and the book title in the subject line.

Daniel Memmert
Editor

Computer Science in Sport

Modeling, Simulation, Data Analysis and Visualization
of Sports-Related Data



Springer

Editor

Daniel Memmert
Institute of Exercise Training and Sport Informatics
German Sport University
Köln, Nordrhein-Westfalen, Germany

ISBN 978-3-662-68312-5 ISBN 978-3-662-68313-2 (eBook)
<https://doi.org/10.1007/978-3-662-68313-2>

Translation from the German language edition: “Sportinformatik” by Daniel Memmert, © Springer-Verlag GmbH 2023. Published by Springer. All Rights Reserved.

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer-Verlag GmbH, DE, part of Springer Nature 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Editorial Contact: Ken Kissinger

This Springer imprint is published by the registered company Springer-Verlag GmbH, DE, part of Springer Nature.

The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

Paper in this product is recyclable

This textbook, *Computer Science in Sport* is dedicated to the unforgettable Jürgen Perl (1944–2023), my esteemed colleague and friend. He was the founder and trailblazing pioneer of Computer Science in Sport, both in Germany and internationally.

Prologue

When Springer-Verlag asked me if I wanted to write a textbook *Computer Science in Sport*, I was very sure from the beginning that I could not do it alone. In recent years, sports informatics has grown extremely, mainly because more and newer data is available. A central building block is certainly the field of game analysis as a performance diagnostic method with which one can conduct systematic analyses in competition, and whose development and dissemination already took its origin in the last millennium. Starting from national to international sports science congresses, central topics of game analysis are currently discussed all over the world (Memmert, 2022).

This textbook wants to do adequate justice to the now broad diversity of sports informatics, in which 33 authors report from their special field and concisely summarize the latest findings. The textbook is divided into four main chapters: data sets, model building, simulation, and data analysis. In addition to a background on programming languages and visualization, it is framed by history and an outlook. It is important to me that the textbook follows a consistent structure within each chapter. Therefore, I am very grateful that all colleagues have used the following structure for their chapters, which hopefully makes it easy for students to find their way through the chapters in a targeted manner. After the four core messages, each chapter is introduced with an example from sports. This is followed by a description of the background of the topic together with a definition. Finally, relevant applications or areas of use are outlined, which are concluded by the detailed presentation of a prominent study that is fundamental to this area in a study box. Before the comprehensive bibliography, additional questions are offered to the students regarding the content of the chapter. The detailed index-word index will hopefully further support the understanding and penetration of the sometimes-complex topics.

I think that sports informatics in German sports science has impressively shown that it has successfully caught up with the parent science of computer science in recent years. This is exemplified by a look at the third-party funding obtained by German sports scientists from the German Research Foundation (DFG) in the review board for computer science (cf. Appendix). Of course, this presentation does not claim to be complete, and some projects are also thematically assigned to sports technology, which is closely related to sports informatics. What they have in common, however, is that scientists from the field of sports informatics and/or sports technology have received competitive projects from the DFG's "computer science funds", in which the guidelines and standards of computer science are applied.

Designing a textbook requires a good mix of intrinsically motivated and sometimes hard work, which a large number of people have done excellently. I would like to take this opportunity to express my sincere thanks to these people. First and foremost, we would like to thank the authors of the book chapters for their willingness to contribute their expertise to our textbook, for their participation in our

internal peer review process, and for their constant desire to improve. Many thanks for the always good and friendly cooperation.

I would also like to thank Ms. Erika Graf for her constant and careful supervision of the book, and her many comments and advice. In addition, I would like to thank our student assistants Klara Rinne, Tara Coulson, and David Brinkjans for taking a critical student perspective, for their constructive feedback to the authors, and for numerous contributions to smooth out rough edges.

Finally, my great thanks go to the constant, very friendly, and always extremely competent support of our book project by the staff of Springer-Verlag. First and foremost, I would like to mention Ken Kissinger (Program Planning), who has put a lot of time, commitment, and energy into the book. This cannot be taken for granted, including his speed in the process! Without his expertise, it would not have come about in this way, and for that, I thank him most sincerely. Regarding the cooperation on the part of Springer-Verlag, I would also like to mention Meike Barth and Anja Herzer (project management), who accompanied the book project very successfully up to the production handover, many thanks for this. I would also like to thank everyone else involved in production (copy-editing, typesetting) for their professional cooperation during the production process.

I hope you enjoy reading this book and that you gain a great deal of knowledge from it.

Daniel Memmert

Cologne, Germany

Contents

I History

1	History	3
	<i>Martin Lames</i>	
1.1	Introduction	4
1.2	The Institutional Constitution of Sports Informatics	5
1.2.1	The Pre-institutional Phase (Before 1995).....	5
1.2.2	The Phase of the dvs Section Sports Informatics (1995–2003).....	6
1.2.3	The Phase of IACSS (2003–2019).....	7
1.2.4	The Institutional Integration Phase of Informatics Working Groups (from 2019).....	7
	References.....	8

II Data

2	Artificial Data	13
	<i>Fabian Wunderlich</i>	
2.1	Example Sport	14
2.2	Background	15
2.2.1	Limits of Real-World Data.....	15
2.2.2	The Idea of Artificial Data.....	15
2.2.3	Random Numbers and Monte Carlo Simulation.....	16
2.2.4	Advantages and Disadvantages of Artificial Data Sets.....	16
2.3	Applications	17
	References.....	19
3	Text Data	21
	<i>Otto Kolbinger</i>	
3.1	Introduction	22
3.2	Applications	23
3.2.1	Evaluation of Technological Officiating Aids.....	23
3.2.2	Match Predictions.....	24
3.2.3	Talent Scouting.....	25
	References.....	26
4	Video Data	27
	<i>Eric Müller-Budack, Wolfgang Gritz, and Ralph Ewerth</i>	
4.1	Example Sport	28
4.2	Background	29
4.3	Basics and Definition	30
4.4	Applications	31
	References.....	33

5	Event Data	35
	<i>Marc Garnica Caparrós</i>	
5.1	Example Sport	36
5.2	Background	37
5.3	Application	38
5.3.1	Event Data to Extend Box Score Statistics.....	38
5.3.2	Event Data to Value in-Game Actions and Player Impact.....	39
5.3.3	Event Data to Understand Player Interactions.....	39
	References.....	40
6	Position Data	43
	<i>Daniel Memmert</i>	
6.1	Example Sport	44
6.2	Background	45
6.3	Applications	46
	References.....	47
7	Online Data	49
	<i>Christoph Breuer</i>	
7.1	Example Sport	50
7.2	Background	51
7.3	Application	52
	References.....	54
III	Modeling	
8	Modeling	57
	<i>Jürgen Perl and Daniel Memmert</i>	
8.1	Example Sport	58
8.2	Background	60
8.3	Application	62
	References.....	63
9	Predictive Models	65
	<i>Fabian Wunderlich</i>	
9.1	Example Sport	66
9.2	Background	67
9.2.1	Looking into the Future.....	67
9.2.2	Predictive Models in Sports.....	67
9.2.3	Creation of Predictive Models.....	68
9.2.4	Exemplary Methods.....	69
9.3	Applications	70
	References.....	71

10	Physiological Modeling	73
	<i>Manuel Bassek</i>	
10.1	Example Sport	74
10.2	Background	75
10.3	Applications	76
	References.....	78

IV Simulation

11	Simulation	81
	<i>Jürgen Perl and Daniel Memmert</i>	
11.1	Example Sport	82
11.2	Background	83
11.3	Applications	86
	References.....	88
12	Metabolic Simulation	89
	<i>Dietmar Saupe</i>	
12.1	Example Sport	90
12.2	Background	91
12.3	Applications	92
	References.....	97
13	Simulation of Physiological Adaptation Processes	99
	<i>Mark Pfeiffer and Stefan Endler</i>	
13.1	Example Sport	100
13.2	Background	101
13.3	Applications	103
	References.....	105

V Programming Languages

14	An Introduction to the Programming Language R for Beginners	109
	<i>Robert Rein</i>	
14.1	History and Philosophy	110
14.2	Concept and Programming Paradigms	111
14.3	Resources on R	112
14.4	R Community and Packages	112
14.5	Introduction to Working with R	113
14.6	An Example Workflow in R	116
	References.....	123

15	Python	125
	<i>Maximilian Klemm</i>	
15.1	Example Sport	126
15.2	Background	127
15.3	Applications	129
	References.....	130
VI	Data Analysis	
16	Logistic Regression	135
	<i>Ashwin Phatak</i>	
16.1	Example Sport	136
16.2	Background	137
16.3	Application	138
	References.....	140
17	Time Series Data Mining	141
	<i>Rumena Komitova and Daniel Memmert</i>	
17.1	Example Sport	142
17.2	Background	143
17.3	Applications	144
17.3.1	Tasks in Time Series Data Mining.....	144
17.3.2	Time Series Data Mining in Medicine	145
17.3.3	Time Series Data Mining in Sports.....	145
	References.....	147
18	Process Mining	149
	<i>Marc Garnica Caparrós</i>	
18.1	Example Sport	150
18.2	Background	151
18.3	Application	153
18.3.1	Process Mining in Healthcare	153
18.3.2	Process Mining in Education.....	153
18.3.3	Process Mining in Soccer.....	153
	References.....	154
19	Networks Centrality	157
	<i>João Paulo Ramos, Rui Jorge Lopes, Duarte Araújo, and Pedro Passos</i>	
19.1	A Network Science in Football	158
19.2	Background	159
19.3	Applications	162
	References.....	166

20	Artificial Neural Networks	169
	<i>Markus Tilp</i>	
20.1	Example Sport	170
20.2	Background	171
20.3	Applications	172
	References.....	176
21	Deep Neural Networks	177
	<i>Dominik Raabe</i>	
21.1	Example Sport	178
21.2	Background	179
21.3	Applications	180
	References.....	183
22	Convolutional Neural Networks	185
	<i>Yannick Rudolph and Ulf Brefeld</i>	
22.1	Example Sport	186
22.2	Background	187
22.3	Applications	189
	References.....	192
23	Transfer Learning	193
	<i>Henrik Biermann</i>	
23.1	Example Sport	194
23.2	Background	195
23.3	Applications	196
	References.....	199
24	Random Forest	201
	<i>Justus Schlenger</i>	
24.1	Example Sport	202
24.2	Background	203
24.3	Applications	204
	References.....	207
25	Statistical Learning for the Modeling of Soccer Matches	209
	<i>Gunther Schauberger and Andreas Groll</i>	
25.1	Example Sport	210
25.2	Background	211
25.3	Applications	212
	References.....	214

26 **Open-Set Recognition**..... 217
Ricardo da Silva Torres

26.1 **Example Sport**..... 218

26.2 **Background**..... 218

26.3 **Applications**..... 220

References..... 221

VII Visualization

27 **Visualization: Basics and Concepts**..... 225
Daniel Link

27.1 **Example Sport**..... 226

27.2 **Background**..... 226

27.3 **Applications**..... 227

References..... 231

VIII Outlook

28 **Outlook**..... 235
Arnold Baca

28.1 **Trends**..... 236

28.2 **Sensors**..... 236

28.3 **Wearables und Intelligent Systems**..... 237

28.4 **Big Data and Cloud**..... 238

28.5 **Machine Learning and Computer Vision**..... 239

28.6 **Virtual und Augmented Reality and Robotics**..... 239

28.7 **Data Protection and Data Misuse**..... 240

References..... 240

Supplementary Information

Appendix. Third-Party Funds Competitively Acquired by German
Sports Scientists from the German Research Foundation (DFG)
in the Review Board for Computer Science..... 244

Index..... 247

Contributors

Duarte Araújo CIPER, Faculdade de Motricidade Humana, Universidade de Lisboa, Cruz Quebrada-Dafundo, Portugal

Arnold Baca Center for Sports Science and University, Sports University of Vienna, Vienna, Austria

Manuel Bassek Institute for Exercise Training and Sports Informatics, German Sports University Cologne, Cologne, Germany

Henrik Biermann Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Cologne, Germany

Ulf Brefeld Leuphana Universität Lüneburg, Lüneburg, Germany

Christoph Breuer German Sport University Cologne, Institute of Sport Economics and Sport Management, Cologne, Germany

Marc Garnica Caparrós Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Cologne, Germany

Stefan Endler Institute of Computer Science, Mainz, Germany

Ralph Ewerth TIB—Leibniz Information Centre for Science and Technology, Hannover, Germany

Wolfgang Gritz TIB—Leibniz Information Centre for Science and Technology, Hannover, Germany

L3S Research Center, Leibniz University Hannover, Hannover, Germany

Andreas Groll Department of Statistics, Statistical Methods for Big Data, TU Dortmund University, Dortmund, Germany

Maximilian Klemp Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Cologne, Germany

Otto Kolbinger TUM School of Medicine and Health, Technical University of Munich, Munich, Germany

Rumena Komitova Institute of Exercise Training and Sport Informatics, German Sports University Cologne, Cologne, Germany

Martin Lames Faculty of Sport and Health Sciences, Technical University of Munich, Munich, Germany

Daniel Link Technical University Munich, Munich, Germany

Rui Jorge Lopes Instituto de Telecomunicações, Iscte, Lisbon, Portugal

Daniel Memmert Institute of Exercise Training and Sport Informatics, German Sports University Cologne, Cologne, Germany

Eric Müller-Budack TIB—Leibniz Information Centre for Science and Technology, Hannover, Germany

L3S Research Center, Leibniz University Hannover, Hannover, Germany

Pedro Passos CIPER, Faculdade de Motricidade Humana, Universidade de Lisboa, Cruz Quebrada-Dafundo, Portugal

Jürgen Perl Institute of Computer Science, FB 08, University of Mainz, Mainz, Germany
Johannes Gutenberg University Mainz, Mainz, Germany

Maerk Pfeiffer Mainz, Germany

Ashwin Phatak Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Cologne, Germany

Dominik Raabe Cologne, Germany

João Paulo Ramos CIDEFES, Universidade Lusófona, Lisbon, Portugal

Robert Rein Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Cologne, Germany

Yannick Rudolph Hamburg, Germany

Dietmar Saupe University of Konstanz, Konstanz, Germany

Gunther Schauburger School of Medicine and Health, Technical University of Munich, Munich, Germany

Justus Schlenger Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Cologne, Germany

Ricardo da Silva Torres Department of ICT and Natural Sciences, NTNU–Norwegian University of Science and Technology, Ålesund, Norway

Wageningen Data Competence Center, Wageningen University and Research, Wageningen, Netherlands

Markus Tilp Institute of Human Movement Science, Sport and Health, University of Graz, Graz, Austria

Fabian Wunderlich Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Cologne, Germany

History

Contents

Chapter 1	History – 3
	<i>Martin Lames</i>



History

Martin Lames

Contents

- 1.1 Introduction – 4**
- 1.2 The Institutional Constitution of Sports Informatics – 5**
 - 1.2.1 The Pre-institutional Phase (Before 1995) – 5
 - 1.2.2 The Phase of the dvs Section Sports Informatics (1995–2003) – 6
 - 1.2.3 The Phase of IACSS (2003–2019) – 7
 - 1.2.4 The Institutional Integration Phase of Informatics Working Groups (from 2019) – 7
- References – 8**

This chapter was translated by Erika Graf and final approved by Martin Lames.

1

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Although rather late compared to other hyphenated computer sciences, sports informatics can show steady progress toward institutionalization as a scientific discipline.
- Stages of this path are free working groups, the dvs section sports informatics, the IACSS, and finally organized bi-directional cooperation in the sense of the sports informatics definition of Link and Lames.
- The future of the discipline depends on whether it will be possible to generate win-win cooperations based on which sustainable interdisciplinary projects can be realized.

1.1 Introduction

If one wants to trace the development of sports informatics as a scientific discipline, one is well advised to first make sure of its subject matter. A generally accepted definition comes from Link and Lames (2015):

Definition

The scientific discipline of "sports informatics" is understood to be multi- and interdisciplinary research programs that contain sports science and informatics components. Its subject is the application of tools, methods and paradigms from computer science to questions of sports science as well as the integration of sports science knowledge into computer science.

1.2 The Institutional Constitution of Sports Informatics

The institutional constitution of sports informatics ¹ as a scientific discipline can be divided into four phases: (1) the pre-institutional phase, (2) the phase of the dvs section Sport Informatics, (3) the phase of the International Association for Computer Science in Sports, and (4) the institutional integration of informatics working groups.

1.2.1 The Pre-institutional Phase (Before 1995)

Characteristic of computer science since its origins, which can be equated with the advent of electronic calculating machines in the middle of the last century, is that besides the core computer science areas of theoretical and technical computer science, the application of these new methods understandably immediately triggered a whole range of scientific activities. One has to differentiate between applied computer science, which established itself within computer science, and the so-called “hyphenated computer sciences”, which (not always to the delight of the “core computer scientists”) are at home in the respective applying science, such as medical computer science or business computer science.

These institutionalizations, which in part brought about “real” interdisciplinary sciences, i.e. those that stand “inter”, i.e. between the two original sciences (Heckhausen, 1986; Willimczik, 1985), were still a long time coming in sports science. Until 1995, there was “only” cooperation in terms of content in research programs with informatics and sports science components, but there were quite a number of them since the ability of informatics to contribute to questions of sports science - as in many other areas - is directly given.

It is therefore somewhat curious that the term “sports informatics” was coined at a very early stage: In 1976, a congress volume entitled “Creative Sports Informatics” was published by Recla and Timmer (1976), which reported on a conference of the “International Association for Sports Information (IASI)” in Graz in 1975. Here, the main interest was in the capabilities of informatics tools to capture, store and flexibly retrieve information from sports, such as information about Olympic participants, which was made available to the press online for the first time in Munich in 1972.

Very early uses of computer science in sports are the computerized game observation systems of the sports educators (sic!) Hagedorn et al. (1980a) in basketball and Brettschneider Allendorf and Brettschneider (1976) in volleyball. Here, conceptually and technologically high-quality work was presented, which unfortu-

1 The phases of the institutional constitution of sports informatics proposed here are by no means identical with the existence and significance of the individual institutions named, but rather represent here significant stages on the path of the institutionalization of sports informatics.

nately was not noticed in the English-speaking world. Interesting are also the studies of groups of computer scientists who either searched for fields of application in sports, e.g. the diploma thesis of Elisabeth André et al. (1988), who later received the Leibniz Prize, about automatic annotation in soccer, or computer scientists who create computer applications out of their enthusiasm for sports, e.g. the TOTO system by Bolch and Cerny (1990), which is based on the Elo system in chess.

The work of Jürgen Perl was to become of decisive importance for the development of the discipline. Together with his colleagues Wolf Miethling and Günter Hagedorn from Paderborn, Perl had already made numerous contributions to sports informatics (Miethling & Perl, 1981; Hagedorn et al., 1980b). At the University of Mainz, he put a scientific focus on sports informatics starting in 1985. Significant contributions of the Mainz group were competition monitoring systems in various sports, each of which was on the cutting edge of information technology. Important on the way to the institutionalization of sports informatics was a workshop series “Sport & Informatik”, which took place regularly from 1989 on.

1.2.2 The Phase of the dvs Section Sports Informatics (1995–2003)

The German Association for Sports Science (dvs) is the scientific organization for academic German sports science. It is organized in sections corresponding to sports science disciplines such as sports education or training science, and commissions representing cross-sectional topics of enduring scientific interest such as sports science contributions to individual sports. Thus, the idea of institutionalizing sports informatics as a section within the framework of the dvs was obvious.

However, achieving this goal was by no means a trivial act, as there were fears within the dvs that the organization would be overstretched by the establishment of many subunits, which furthermore corresponded to the hitherto extremely influential position of Grupe’s “Integrative Sports Science” (Krüger, 2015). Furthermore, some groups, such as the working group “Media in Sport”, saw their claim to representation in terms of content threatened by a sports informatics section.

In addition to numerous discussions with dvs functionaries and colleagues with an affinity for the subject, and of course, the reference to an existing, and indeed interdisciplinary, community that had manifested itself at the regular workshops, an article was placed in the journal “Leistungssport” (Competitive Sports) in which the subject area was presented to the German public under the title “Sport Informatics: Gegenstandsbereich und Perspektiven einer sportwissenschaftlichen Teildisziplin” (Perl & Lames, 1995). There, the potential of sports informatics for sports science was explained, especially concerning its connectivity to topics of other sections and its ability to contribute to the support of elite sports.

Since the approval of the general meeting at the dvs-Hochschultag 1995 in Frankfurt to found the section, the “Workshop Sport & Informatik” now operated

as a meeting of the “dvs section Sportinformatik”. The section successfully asserts itself in the market of scientific organizations and organizes with the speakers Perl (1996–2002), Wiemeyer (2002–2012), Lames (2012–2018), and Link (from 2018) its section conferences in even years at the different centers of the discipline in Germany (2014: Vienna). The 2018 conference in Munich will be held for the first time with a section on “Sports Informatics and Sports Technology”, thus future-proofing the corresponding development in the two sciences.

1.2.3 The Phase of IACSS (2003–2019)

For Jürgen Perl, it was clear very early on that the institutionalization of sports informatics could not stop at a national sports science section, but that the international stage also had to be “played”. Strategically, the same path was followed on the national level. The first international congresses in Cologne (1997), Vienna (1999), and Cardiff (2001) gathered a critical mass of international scientists, which then made it possible to launch the IACSS (International Association of Computer Science in Sport) at the conference in Barcelona (2003).

Subsequently, the effort was not to limit the IACSS to Europe, which would not have corresponded to the claim of an international scientific organization. Meetings in Canada, China, Australia, and Brazil testify to the success of these efforts. The establishment of national associations for sports informatics, as the umbrella organization of the IACSS, was only successful in a few countries (e.g. Germany, Austria, Turkey, Russia, and China). Mainly probably because initiatives to found a national association, if they are essentially only based on the initiative of individual research personalities, did not prove to be sustainable. The recognition of the IACSS as a member of the ICSSPE (International Council for Sport Science and Physical Education), the umbrella organization for sports science associations, was an important step towards institutional consolidation.

At a very early stage, under the editorship of Arnold Baca (from 2002), a journal, the IJCSS (International Journal of Computer Science in Sport) was established, which can be seen as another important characteristic of the degree of institutionalization of science. Since 2016, the journal has been published as an open-access journal together with DeGruyter Verlag.

The respective IACSS presidents Perl (2003–2007), Baca (2007–2013), Lames (2013–2022), and Zhang (from 2022) have succeeded in establishing a globally recognized and active organization with regular meetings and publication activities.

1.2.4 The Institutional Integration Phase of Informatics Working Groups (from 2019)

At the IACSS General Assembly in Moscow in 2019, it was decided that the IACSS should move more in the direction of computer science and integrate existing working groups there that deal with the topic of sports. Successfully in this direc-

tion so far have been in contact with the MLSA (Machine Learning in Sports Analytics) group around Jesse Davis, Jan van Haaren, Albrecht Zimmermann, and Ulf Brefeld. This group has been organizing either its own workshops or satellite workshops at major computer science conferences since 2013, most recently in Grenoble in 2022; before that, a virtual workshop was organized in 2021 and—just like the 7 workshops before it—documented in proceedings (Brefeld et al., 2022). Collaboration with other, comparable groups, for example from the field of computer vision, is still pending.

An important instrument for the integration of computer scientists and sports scientists are seminars in Schloss Dagstuhl, an international conference center for computer science. Of the five seminars on sports science topics hosted with IACSS participation to date, the most recent was held jointly with MLSA in October 2021 on “Machine Learning in Sports” (organizers: Brefeld, Davis, Lames, Little). In the future, attendance at each other’s meetings will be required, and ideally, we will succeed in establishing project groups that also represent sports science and computer science in their ranks in terms of personnel.

Institutional integration is all the more important because the integration of two disciplines must not be viewed naively. For example, sports scientists often rely uncritically on the answers provided by computer science without being able to question its basic methodological assumptions. On the other hand, computer scientists often use the meanwhile good data situation in the attractive application field of sports (professional soccer) only as a showcase for their original basic scientific questions. These problems could at least be reduced with suitably designed institutionalized cooperation and thus offer the enormous potential of sports informatics opportunities for development.

🔍 Questions for the Students

1. How does the development of sports informatics compare to that of other sports science disciplines?
2. Why is it important for project groups in sports informatics to be interdisciplinary, and what role do science organizations play in this?

References

-
- Allendorf, O. & Brettschneider, W.-D. (1976). Leistungsdatenerfassung und -auswertung im Sportspiel mit Hilfe des computergesteuerten optischen Lesestifts. In R. Andresen & G. Hagedorn (Hrsg.), *Zur Sportspielforschung* (Band 1: Theorie und Praxis der Sportspiele; S. 106–116). Bartels & Wernitz.
- André, E., Herzog, G., & Rist, Th. (1988). On the simultaneous interpretation of real world image sequences and their natural language description: The system SOCCER. Proceedings of the eighth conference on artificial intelligence, Munich, August 1-5, 1988 (pp. 449–454).
- Bolch, G. & Cerny, M. (1990). TOTO: Ein Programmsystem zur Analyse und Prognose der Ergebnisse von Mannschaftsspielen mit Hilfe eines IBM PC. In J. Perl (Hrsg.), *Sport und Informatik*. 1. Workshop Sport & Informatik, Hochheim, 26.-27.4.89 (S. 25–35). Hofmann.
- Brefeld, U., Davis, J., Van Haaren, J., & Zimmermann, A. (Eds.). (2022). Machine learning and data mining for sports analytics—*8th International workshop MLSA 2021, virtual event, September 13,*

- 2021 (Springer Conference Proceedings, Communications in Computer and Information Science, Vol. 1571). Springer.
- Hagedorn, G., Ehrich, D., & Schmidt, G. (1980a). Computerunterstützte Spielanalyse im Basketball. *Leistungssport*, 10(5), 363–372.
- Hagedorn, G., Lorenz, H., & Meseck, U. (1980b). Die Verteilung spieltypischer Aktivitäten im Basketball. *Leistungssport*, 11(6), 442–449.
- Heckhausen, H. (1986). Interdisziplinäre Forschung zwischen Intra-, Multi- und Chimären-Disziplinarität. In Zentrum für interdisziplinäre Forschung der Universität Bielefeld (ZIF) (Hrsg.), *Jahresbericht 1985/86* (S. 29–40). ZIF.
- Krüger, M. (2015). Ommo Grupe und seine Vision des Sports. *Sportwissenschaft*, 45, 55–56.
- Link, D., & Lames, M. (2015). An introduction to sport informatics. In A. Baca (Ed.), *Computer Science in Sport – Research and Practice* (pp. 1–17). Routledge.
- Miethling, W.-D., & Perl, J. (1981). *Computerunterstützte Sportspielanalyse*. Czwalina.
- Perl, J., & Lames, M. (1995). Sportinformatik: Gegenstandsbereich und Perspektiven einer sportwissenschaftlichen Teildisziplin. *Leistungssport*, 25(3), 26–30.
- Recla, J., & Timmer, R. (Eds.). (1976). *Kreative Sportinformatik*. Hofmann.
- Willimczik, K. (1985). Interdisziplinäre Sportwissenschaft—Forderungen an ein erstarrtes Konzept. *Sportwissenschaft*, 15, 9–32.



Data

Contents

- Chapter 2** **Artificial Data – 13**
Fabian Wunderlich
- Chapter 3** **Text Data – 21**
Otto Kolbinger
- Chapter 4** **Video Data – 27**
*Eric Müller-Budack, Wolfgang Gritz,
and Ralph Ewerth*
- Chapter 5** **Event Data – 35**
Marc Garnica Caparrós
- Chapter 6** **Position Data – 43**
Daniel Memmert
- Chapter 7** **Online Data – 49**
Christoph Breuer



Artificial Data

Fabian Wunderlich

Contents

- 2.1 Example Sport – 14**
- 2.2 Background – 15**
 - 2.2.1 Limits of Real-World Data – 15
 - 2.2.2 The Idea of Artificial Data – 15
 - 2.2.3 Random Numbers and Monte Carlo Simulation – 16
 - 2.2.4 Advantages and Disadvantages of Artificial Data Sets – 16
- 2.3 Applications – 17**
- References – 19**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- The availability of real-world data sets from the sports domain can be severely limited, especially with regard to aspects like size, consistency, or quality.
- So-called random number generators can be used to simulate random numbers that satisfy certain predefined characteristics.
- By means of these random numbers, complex processes can be replicated, their outcomes can be simulated and artificial data sets can be generated without actually having to observe these processes in reality.
- Some advantages of artificial datasets are the almost unlimited availability and the precise knowledge about the underlying processes.
- The main disadvantage is that the value of artificial data is highly dependent on whether the assumptions made during its creation accurately represent reality.

2.1 Example Sport

When using sports as an application for computer science methods, we aim at achieving improved classifications, making better decisions, detecting hidden patterns in the data, and most importantly, at better understanding the processes in sports. How does the serve ability of a tennis player impact his or her probability of winning the match? Which defensive tactic of a soccer team increases its probability of success? How does training load in American football affect the players' risk of injury? From a data perspective, the complexity of such questions can be increased by the limited availability, consistency, and quality of real-world data. Moreover, in professional sports, researchers face the problem of being unable to experimentally manipulate real-world competitions. By means of simulation and artificial data, however, such questions can be investigated without having to rely on experimentally collected or observed real-world data.

2.2 Background

2.2.1 Limits of Real-World Data

A large part of this book discusses the analysis of real-world data sets. Using these data sets is the natural approach, as they represent the results of the complex real-world processes, that we want to analyse by adopting methods from computer science. However, it should be kept in mind that even real data sets can be subject to severe problems. In particular, a major challenge is that real-world data may be limited concerning various aspects such as size, consistency, and competition rules. In many applications, the sample size does not present a problem, for example, when analysing a large number of match results (Angelini & de Angelis, 2019; Kovalchik, 2016) or goals (Wunderlich et al., 2021a). However, natural limits exist due to the rarity of certain events, such as World Cup matches (Armatas et al., 2007; Delgado-Bordonau et al., 2013). Moreover, for data with a high granularity such as positional data, data availability is still significantly limited and it is still common that studies are based on 50 or less matches (Clemente et al., 2014; Klemp et al., 2021).

The consistency of the data can be affected by internal and external factors, as professional sports can be subject to inconsistent rules, rule changes, or social influences. For example, male tennis players competing in a best of five sets format in Grand Slam tournaments while playing best of three sets in ATP Tour matches (see Clarke & Dyte, 2000), the awarding of three instead of two points for a win in soccer (Riedl et al., 2015), the adjustment of the three-point line in basketball (Strumbelj et al., 2013), or possible effects of ghost games during the COVID-19 pandemic on home advantage (Wunderlich et al., 2021b). Further data-limiting issues include the completeness and veracity of available data.

2.2.2 The Idea of Artificial Data

The idea of artificial data, in a way, can be compared to the idea of hypothesis testing in statistics. When performing such tests, we assume some null hypothesis to be true, and based on this, the distribution of the possible outcomes of an experiment is calculated (based on the hypothesis being true and taking randomness into account). Once the experiment has been performed, this theoretical distribution helps to understand how well the experimental data match the given hypothesis. To put it simple, it helps to decide whether the data found tend to be in favour of the null hypothesis or against it.

While probability distributions in hypothesis testing can be stated explicitly, artificial data usually applies to situations, which are too complex to be described explicitly through mathematical formulas with reasonable effort. Instead, in a first step, the underlying process is modelled mathematically by describing the system-

atic characteristics of the process. Using random numbers and so-called Monte Carlo simulation, the random aspects can then be added in a second step. Thus, the probability of intermediate results or outcomes of the process can be estimated.

2.2.3 Random Numbers and Monte Carlo Simulation

Random numbers are machine-generated numbers that correspond to a certain predetermined probability distribution (James, 1990). In this way, for example, 10,000 numbers whose probability of occurrence corresponds to a Poisson distribution with a predefined mean, but which otherwise have a random character, can be generated. If the characteristics of a simple process are known, the random results of this process can be generated repeatedly. This can be illustrated by the example of goals in soccer, whose number is well approximated by a Poisson distribution (Karlis & Ntzoufras, 2003). Thus, by drawing random numbers from a Poisson distribution, realistic values for the number of goals in a large set of games can be artificially simulated without being in need to observe such a large set of games. A simulation of a more complicated process usually requires the generation of a wide variety of variables, which can additionally influence and interact with each other. The data sets created by mathematical modelling of the process and generating random numbers are denoted as artificial data. The process of generation is synonymously referred to as simulation or Monte Carlo simulation (Harrison, 2010).

2.2.4 Advantages and Disadvantages of Artificial Data Sets

Several of the limitations of real-world data mentioned above can be overcome by artificial data, as it enables the researcher to simulate data with an almost unlimited size, as well as with complete consistency and quality. At this point, it shall be mentioned that large sample sizes are particularly important in situations involving a high degree of randomness, which demonstrably applies to the domain of sports games (Ben-Naim et al., 2006; Brechot & Flepp, 2020; Lames, 2018; Wunderlich et al., 2021a). Artificial data, which can be generated faster and in larger quantities than the available real-world data in certain domains, can be a valuable tool. Another advantage is that, in contrast to real-world processes, the relationships inherent in the artificial data can be intentionally controlled and varied to understand the influence of different variables on the process outcome.

The most important disadvantage of artificial data is the limited transferability to real-world processes. Obviously, the results of a simulation model are directly dependent on the modelling of the process and the assumptions made about the probability distributions of the individual variables. In this respect, the value of the artificial data and conclusions drawn from them depend strongly on whether the assumptions made during their creation precisely represent the processes present in reality. Complexity and difficulty to observe a process in reality drives the risk of

using imprecise assumptions in the modelling. The more complex a process, the more the simulation results can be considered theoretical. As a consequence, an interplay of artificial data and real-world data is needed. In a first step, theoretical knowledge about the underlying processes can be found by using artificial data. In a second step, the actual benefit in practice can be proven by transferring these insights to real-world data.

The method of generating artificial data through simulation is almost universally applicable and its areas of application include healthcare (Jahangirian et al., 2012; Zhang, 2018), manufacturing (Mourtzis et al., 2014) or testing of software systems (Misra, 2015), among many other examples. In the domain of sports, simulation and artificial data techniques are used for a wide variety of issues, some of which are discussed in more detail below in this chapter (Bornn et al., 2019; Garnica-Caparrós et al., 2022; Leitner et al., 2010; Memmert et al., 2021; Newton & Aslam, 2009; Štrumbelj & Vračar, 2012; Wunderlich & Memmert, 2020).

Definition

In this chapter, artificial data means data generated by statistical modelling of a process and applying Monte Carlo simulation. These data are usually intended to represent the results of a real-world process in sports and are used to improve our understanding of this process or related processes.

Study Box

In their study, Garnica-Caparrós et al. (2022) present a simulation framework that can be used to generate and analyse artificial data. Predictive models for forecasting sports events serve as an application example. Conceptually, the entire prediction process is replicated by simulating team or player strengths, match outcomes as well as betting odds influenced by bookmaker errors. Existing or novel rating methods and forecasting models can then be validated based on artificial data sets. The advantage of the artificial data sets compared to real-world data sets is, among other things, that they enable a more precise validation of the accuracy and profitability of the forecasting models.

2.3 Applications

► Example 1

The first domain of application focuses on simulation of the outcomes of sports events based on certain assumptions. Usually, both the estimation of the systematic strength of teams or players as well as a modelling of the course of play in the respective sport are required. Under these assumptions, matches or entire competitions can be repeatedly simulated. Based on a large number of such simulation runs, the probability of each of

the possible outcomes is obtained. The study by Newton and Aslam (2009) represents an exemplary case of this approach. Using real data from professional tennis, the authors estimate the systematic strength of tennis players, measured as the probability of winning a point on their serve or return. In addition, they estimate the variation in players' performance, measured as the variability of these probabilities. Using a model that builds upon the rules of tennis involving points, games and sets, the probabilities for different match outcomes can be derived from the player characteristics via Monte Carlo simulation. Even entire tournaments can be simulated by randomly drawing the winners of each match under the given probabilities. Artificial data on tournament outcomes can then be analysed to determine the most likely winners and each player's chances of reaching a particular round. ◀

► Example 2

The second domain of application aims to better understand the impact and interaction of different influencing factors. In this regard, simulation and artificial data can be used to understand the impact of individual variables and their interaction on the outcome of a process and, above all, to avoid incorrect conclusions. An exemplary case is the study by Bornn et al. (2019), which addresses the question of whether workload, measured as the so-called acute chronic workload ratio (ACWR), is a predictor of injury risk. Numerous existing studies had previously suggested that this is true, while the authors of the aforementioned study suspect that other influencing factors may confound this relationship. For this reason, artificial data were simulated under the assumption that injuries depend only on the load of the current training session and not on the ACWR. However, when analysing the resulting data, significant correlations between ACWR and injuries can still be found. Solely based on theoretical considerations and artificial data, the authors were able to prove that the load of the current training session can confound the results, and thus previously found results using real-world data may have been inaccurate or even misinterpreted. ◀

► Example 3

The third use case applies artificial data to overcome two challenges of real complex data sets from sports. First, the specific situation to be analysed is rarely available due to the generally limited number of matches, where complex data is available. Second, the issue that experimental manipulations, which would allow the deliberate creation of such situations, are not possible in professional sports events. Artificial data can enable researchers to conduct meaningful analysis, despite these issues. An exemplary case is a study by Memmert et al. (2021), in which artificial data were used to extend the analysis of positional data in soccer. Through a simulation approach, it is possible to systematically investigate various combinations of formation flexibility of the attacking and defensive team. In particular, it is possible to find out which tactical flexibility of the teams is most promising. Although real-world positional data is used as the basis for the simulation, a pure analysis of positional data from real matches is not sufficient, as it would not be guaranteed that all combinations to be analysed are actually present in these matches. ◀

? Questions for the Students

1. What is Monte Carlo simulation?
2. What are the main advantages and disadvantages of artificial data?

References

- Angelini, G., & de Angelis, L. (2019). Efficiency of online football betting markets. *International Journal of Forecasting*, 35(2), 712–721. <https://doi.org/10.1016/j.ijforecast.2018.07.008>
- Armatas, V., Yiannakos, A., & Sileloglou, P. (2007). Relationship between time and goal scoring in soccer games: Analysis of three world cups. *International Journal of Performance Analysis in Sport*, 7(2), 48–58. <https://doi.org/10.1080/24748668.2007.11868396>
- Ben-Naim, E., Vazquez, F., & Redner, S. (2006). Parity and predictability of competitions. *Journal of Quantitative Analysis in Sports*, 2(4). <https://doi.org/10.2202/1559-0410.1034>
- Bornn, L., Ward, P., & Norman, D. (2019). Training schedule confounds the relationship between acute: Chronic workload ratio and injury. Sloansportsconference Com.
- Brechot, M., & Flepp, R. (2020). Dealing with randomness in match outcomes: How to rethink performance evaluation in European Club Football using expected goals. *Journal of Sports Economics*, 21(4), 335–362. <https://doi.org/10.1177/1527002519897962>
- Clarke, S. R., & Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research*, 7(6), 585–594. <https://doi.org/10.1111/j.1475-3995.2000.tb00218.x>
- Clemente, M. F., Martins, F. M. L., Couceiro, S. M., Mendes, S. R., & Figueiredo, A. J. (2014). Inspecting teammates' coverage during attacking plays in a football game: A case study. *International Journal of Performance Analysis in Sport*, 14(2), 384–400. <https://doi.org/10.1080/24748668.2014.11868729>
- Delgado-Bordonau, J. L., Domenech-Monforte, C., Guzmán, J. F., & Méndez-Villanueva, A. (2013). Offensive and defensive team performance: Relation to successful and unsuccessful participation in the 2010 Soccer World Cup. *Journal of Human Sport and Exercise*, 8(4), 894–904. <https://doi.org/10.4100/jhse.2013.84.02>
- Garnica-Caparrós, M., Memmert, D., & Wunderlich, F. (2022). Artificial data in sports forecasting: A simulation framework for analysing predictive models in sports. *Information Systems and e-Business Management*, 551–580. <https://doi.org/10.1007/s10257-022-00560-9>
- Harrison, R. L. (2010). Introduction to Monte Carlo simulation. *AIP Conference Proceedings*, 1204, 17–21. <https://doi.org/10.1063/1.3295638>
- Jahangirian, M., Naseer, A., Stergioulas, L., Young, T., Eldabi, T., Brailsford, S., et al. (2012). Simulation in health-care: Lessons from other sectors. *Operational Research*, 12(1), 45–55. <https://doi.org/10.1007/s12351-010-0089-8>
- James, F. (1990). A review of pseudorandom number generators. *Computer Physics Communications*, 60(3), 329–344.
- Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (the Statistician)*, 52(3), 381–393. <https://doi.org/10.1111/1467-9884.00366>
- Klemp, M., Wunderlich, F., & Memmert, D. (2021). In-play forecasting in football using event and positional data. *Scientific Reports*, 11(1), 24,139. <https://doi.org/10.1038/s41598-021-03157-3>
- Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, 12(3). <https://doi.org/10.1515/jqas-2015-0059>
- Lames, M. (2018). Chance involvement in goal scoring in football—An empirical approach. *German Journal of Exercise and Sport Research*, 48(2), 278–286. <https://doi.org/10.1007/s12662-018-0518-z>
- Leitner, C., Zeileis, A., & Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, 26(3), 471–481. <https://doi.org/10.1016/j.ijforecast.2009.10.001>
- Memmert, D., Imkamp, J., & Perl, J. (2021). Flexible defense succeeds creative attacks!—A simulation approach based on position data in professional football. *Journal of Software Engineering and Applications*, 14(09), 493–504. <https://doi.org/10.4236/jsea.2021.149029>
- Misra, A. (2015). Comparative study of test data generation techniques. *JITS*, 1(2), 1–7.
- Mourtzis, D., Doukas, M., & Bernidaki, D. (2014). Simulation in manufacturing: Review and challenges. *Procedia CIRP*, 25, 213–229. <https://doi.org/10.1016/j.procir.2014.10.032>

- Newton, P. K., & Aslam, K. (2009). Monte Carlo tennis: A stochastic Markov chain model. *Journal of Quantitative Analysis in Sports*, 5(3). <https://doi.org/10.2202/1559-0410.1169>
- Riedl, D., Heuer, A., & Strauss, B. (2015). Why the three-point rule failed to sufficiently reduce the number of draws in soccer: An application of prospect theory. *Journal of Sport & Exercise Psychology*, 37(3), 316–326. <https://doi.org/10.1123/jsep.2015-0018>
- Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting*, 28(2), 532–542. <https://doi.org/10.1016/j.ijforecast.2011.01.004>
- Štrumbelj, E., Vračar, P., Robnik-Šikonja, M., Dežman, B., & Erčulj, F. (2013). A decade of euroleague basketball: An analysis of trends and recent rule change effects. *Journal of Human Kinetics*, 38, 183–189. <https://doi.org/10.2478/hukin-2013-0058>
- Wunderlich, F., & Memmert, D. (2020). Are betting returns a useful measure of accuracy in (sports) forecasting? *International Journal of Forecasting*, 36(2), 713–722. <https://doi.org/10.1016/j.ijforecast.2019.08.009>
- Wunderlich, F., Seck, A., & Memmert, D. (2021a). The influence of randomness on goals in football decreases over time. An empirical analysis of randomness involved in goal scoring in the English Premier League. *Journal of Sports Sciences*, 39(20), 2322–2337. <https://doi.org/10.1080/02640414.2021.1930685>
- Wunderlich, F., Weigelt, M., Rein, R., & Memmert, D. (2021b). How does spectator presence affect football? Home advantage remains in European top-class football matches played without spectators during the COVID-19 pandemic. *PLoS One*, 16(3), e0248590. <https://doi.org/10.1371/journal.pone.0248590>
- Zhang, X. (2018). Application of discrete event simulation in health care: A systematic review. *BMC Health Services Research*, 18(1), 687. <https://doi.org/10.1186/s12913-018-3456-4>



Text Data

Otto Kolbinger

Contents

- 3.1 Introduction – 22**
- 3.2 Applications – 23**
 - 3.2.1 Evaluation of Technological Officiating Aids – 23
 - 3.2.2 Match Predictions – 24
 - 3.2.3 Talent Scouting – 25
- References – 26**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <http://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Text data can contain information about athletes, competitions, and their impact on society that is not contained in other data
- Computer-aided text mining methods allow economic processing and analysis of large (text) datasets
- Text data are currently mainly used to address sociological and economic issues related to sport
- First studies show potential for the systematic analysis of text data also in areas that are primarily concerned with the performance of athletes, such as exercise science and performance analysis.

3.1 Introduction

For a long time, the processing of questions from sports and sports science with informatic methods dealt almost exclusively with numerical data, such as action or position data. However, knowledge about athletes, competitions, and their effects is often also available in textual form, such as countless scouting reports in junior academies of clubs and federations. Nowadays, advances in text recognition and machine learning allow the efficient analysis of large text datasets. Accordingly, so-called "text mining" methods are increasingly used in theory and practice, especially in disciplines that traditionally work a lot with data in text form such as open-ended questionnaires or standardized interviews. In sports science, for example, studies from the field of sports sociology dominate. Several studies have already examined how fans react to political statements by athletes, for example, Frederick et al. (2020) regarding political statements by Megan Rapinoe or Schmidt et al. (2019) regarding protests during the national anthem. Both of the studies listed used social media posts as their data source—another trend in research based on textual data.

Accordingly, the exemplary applications in this chapter show two studies that are based on social media data.

One study deals with the influence of the video assistant referee in soccer, the so-called VAR, on the mood of fans of the English Premier League on Twitter (Kolbinger & Knopp, 2020). Here, text data can be used to systematically investigate the impact of technical innovations on stakeholders. An aspect that has been neglected for the introduction of technological officiating aids for referees. How posts from social media can further be used to predict game outcomes is also demonstrated in this chapter using an example from American football (Schumaker et al., 2017). The fact that this is not the only textual data that can contain valuable information for match prediction is discussed using a study by Beal et al. (2021). Finally, as a last exemplary application area, two papers will demonstrate how text data can be used in talent scouting (Maymin, 2021; Seppa et al., 2017).

Definition

All structured and unstructured text bodies can in principle serve as text data. In the context of sports informatics, this includes, but is by no means limited to, social media posts, interviews, and expert assessments in text form.

3.2 Applications

3.2.1 Evaluation of Technological Officiating Aids

Over the last decades, more and more sports introduced so-called Technological Officiating Aids to support referees (Kolbinger, 2018). While sports practice and academic publications have focused primarily on the technology itself and its impact on decision quality, the influence of these interventions on stakeholders such as fans has been neglected (Kolbinger & Lames, 2017). Here, textual data in the form of social media posts offer an easily accessible way to study precisely this influence. Over 3 billion people worldwide use platforms such as Twitter, Facebook, or Instagram to make their opinions and emotions known on specific topics or events (Kozinets, 2020). This of course includes polarizing topics in sports and the video assistant in soccer can undoubtedly be seen as such (see also Kolbinger 2020).

As a data basis for the study on the influence of the VAR on the sentiment of soccer fans on Twitter, Kolbinger and Knopp (2020) used all tweets of 129 matches of the Premier League season 2019/20 in which the official match hashtag was used (i.e., for example, #LIVMUN for Liverpool FC vs. Manchester United FC). Of these total 643,251 tweets, 58,264, or 9.1%, dealt with the video referee. For these tweets, as well as the rest of the sample, the authors also performed so-called sentiment analysis, to evaluate whether a post expressed an overall rather negative, neutral, or positive sentiment. What was striking was that 76.2% of the tweets about the VAR expressed negative emotions and only 12.3% expressed positive emotions.

In contrast, for all other tweets during the soccer matches considered, 39.4% of the posts were positive and only 31.3% were negative. In addition, Kolbinger and Knopp (2020) looked at how video referee interventions affected average sentiment during soccer matches and were able to show that on average these interventions led to a significant drop in sentiment that lasted over 20 min.

For both deciding whether a tweet referred to the VAR and assessing sentiment, the authors developed and used automatic text classifiers. This means that an algorithm was trained to automatically classify the content of posts into the aforementioned categories. This is a common procedure for analyzing large amounts of text data, and it was also used in many of the following studies. For all these studies, the quality of the text classifier is of paramount importance. A circumstance that, as with other applications of machine learning methods, is unfortunately often neglected. Accordingly, it is important for readers of studies with text data that not only the origin and type of the data are comprehensible, but also the applied classification procedure and its quality (Kolbinger, 2022).

3.2.2 Match Predictions

Another area of application in which the sentiment of social media posts has already been used in a promising way is the prediction of match results. An American research group led by Robert P. Schumaker demonstrated this in one study each on the English Premier League (Schumaker et al., 2016) and the National Football League (NFL—American Football; Schumaker et al., 2017). It is more or less an attempt to use the so-called “Wisdom of the Crowd” to predict the outcome of games. In this subsection, we focus on the NFL study, in which the authors used a very interesting approach. For each game, all tweets about one of the participating teams in the 96 h (4 days) before kickoff served as the data basis. Schumaker et al. (2017) compared how the average sentiment of these posts changed on the last day before the game compared to the 3 days before. They simply predicted the team for which there was a more positive change in sentiment (or a less negative one) as the winner of the game. With this simple method, the authors achieved the same prediction rate as sports betting providers, and they were particularly good at predicting wins of underdogs.

The same pattern is found in a paper by Beal et al. (2021), but using a fundamentally different type of text data. Instead of using many nonspecific short text corpora hoping for swarm intelligence, as Schumaker et al. (2017) did, the text data of Beal et al. (2021) are each single, elaborate texts on a predefined topic. In more detail, preliminary reports on Premier League matches from an English daily newspaper. Here, in contrast to the previous studies, no classifications of the text (except for the assignment of each sentence to a team) were made. Again, this approach specifically predicted unexpected outcomes better than models based on numerical data. Thus, it seems that there is information contained in the text that cannot (at least not yet) be represented by numerical data like previous results or so-called key performance indicators.

3.2.3 Talent Scouting

“[Kobe knows] the game of basketball and what needs to be done to win”—this excerpt from Jason Sean Fuiman’s scouting report on Kobe Bryant, who sadly passed away much too soon, proved to be very true (Sumsky, 2020). Interestingly, Fuiman refers here to performance characteristics of Kobe Bryant that have not been - and may never be—mapped via numerical data. Such scouting reports exist in countless clubs and associations for a countless number of athletes. Text-mining techniques can harness them in an economical way, as it is demonstrated by two studies which each attempted to predict expected performance in a professional league based on commercial scouting reports. Both used very different approaches to do so.

Seppa et al. (2017) combined sentiment analysis with a so-called lexicon-based categorization of text data to evaluate scouting reports in ice hockey. This means that they tried to assign each sentence or paragraph of a scouting report to a category via certain keywords or word strings. For example, a sentence was assigned to the category “effort” if words such as “effort” itself or “lazy” or word chains such as “needs to compete harder” appeared. In conjunction with an analysis of the sentiment of the corresponding passage, they then tried to classify the players for each of these categories, for example as players with “poor effort” or “good puck skills”. In this way, Seppa et al. (2017) were able to predict the assist and goal rates of players in the professional leagues better compared to models based on the assists and goals scored in youth leagues. Predictions were even better when the two data types were combined. However, it must be noted that the prediction quality for both methods used and their combination was not very high. However, the pattern that the reports were able to improve the prediction quality is very interesting.

In a study on basketball, Maymin (2021) describes that a model he developed would have outperformed the draft performance of 29 of the 30 NBA teams. In addition to game statistics, the model included scouting reports, which contained scores for specific skills as well as pre-structured text data. The text bodies were divided into “Strengths,” “Weaknesses,” “Overall,” and “Notes.” Because Maymin (2021) was primarily concerned with comparing his model to the draft performance of NBA teams, he did not go into detail about the contributions of the individual components of his model. However, it is already apparent from the summary overview of the importance of the components that the content and, in particular, both length and sentiment of the individual categories made a predictive contribution that was similar to that of individual match statistics.

Accordingly, the state of research on the usability of text data for talent diagnostics is by no means satisfactory, especially since the few studies to date have focused on commercial scouting reports and exclusively on the sports system in North America. However, the initial results can certainly be considered promising and suggest that text data represent information that is not (yet) available in other data.

? Questions for the Students

1. name two methods to automatically classify text data.
2. describe with two examples how to use social media posts in sports science.

3

References

- Beal, R., Middleton, S. E., Norman, T. J., & Ramchurn, S. D. (2021). Combining machine learning and human experts to predict match outcomes in football: A baseline model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17), 15,447–15,451.
- Frederick, E. L., Pegoraro, A., & Schmidt, S. (2020). “I’m not going to the f***ing white house”: Twitter users react to Donald Trump and Megan Rapinoe. *Communication & Sport, in press*, 10, 1210–1228. <https://doi.org/10.1177/2167479520950778>
- Kolbinger, O. (2018). *Innovative technische Hilfsmittel zur Unterstützung von Schiedsrichtern in Spielsportarten als Gegenstand von Evaluationsforschung [Innovative Technological Officiating Aids as object of Evaluative Research]*. Doctoral dissertation, Technical University of Munich.
- Kolbinger, O. (2020). VAR experiments in the Bundesliga. In M. Armenteros, A. J. Benítez, & M. A. Betancor (Eds.), *The use of video technologies in refereeing football and other sports* (pp. 228–245). Routledge.
- Kolbinger, O. (2022). Text mining and performance analysis. In: *International conference on security, privacy, and anonymity in computation, communication, and storage* (pp. 3–8). Springer, Cham.
- Kolbinger, O., & Knopp, M. (2020). Video kills the sentiment—Exploring fans’ reception of the video assistant referee in the English premier league using twitter data. *PLoS One*, 15(12), e0242728. <https://doi.org/10.1371/journal.pone.0242728>
- Kolbinger, O., & Lames, M. (2017). Scientific approaches to technological officiating aids in game sports. *Current Issues in Sport Science*, 2, 001. https://doi.org/10.15203/CISS_2017.001
- Kozinets, R. V. (2020). *Netnography: The essential guide to qualitative social media research*. Sage.
- Maymin, P. (2021). Using scouting reports text to predict NCAA→NBA performance. *Journal of Business Analytics*, 4(1), 40–54. <https://doi.org/10.1080/2573234X.2021.1873077>
- Schmidt, S. H., Frederick, E. L., Pegoraro, A., & Spencer, T. C. (2019). An analysis of Colin Kaepernick, Megan Rapinoe, and the national anthem protests. *Communication & Sport*, 7(5), 653–677. <https://doi.org/10.1177/2167479518793625>
- Schumaker, R. P., Jarmoszko, A. T., & Labeledz, C. S., Jr. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decision Support Systems*, 88, 76–84. <https://doi.org/10.1016/j.dss.2016.05.010>
- Schumaker, R. P., Labeledz, C. S., Jr., Jarmoszko, A. T., & Brown, L. L. (2017). Prediction from regional angst—a study of NFL sentiment in Twitter using technical stock market charting. *Decision Support Systems*, 98, 80–88. <https://doi.org/10.1016/j.dss.2017.04.010>
- Seppa, T., Schuckers, M. E., & Rovito, M. (2017). Text mining of scouting reports as a novel data source for improving NHL draft analytics. In *Ottawa hockey analytics conference* (pp. 1–11).
- Sumsky, A. (2020, July 2). Kobe Bryant’s Scouting Report is Worth the Read. Basketball forever. <https://basketballfor-ever.com/2020/07/02/kobe-bryants-scouting-report-worth-read>.



Video Data

Eric Müller-Budack, Wolfgang Gritz, and Ralph Ewerth

Contents

- 4.1 Example Sport – 28**
- 4.2 Background – 29**
- 4.3 Basics and Definition – 30**
- 4.4 Applications – 31**
- References – 33**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Video data capture actions and poses of athletes as well as movements.
- With the help of AI-based approaches, video recordings can be automatically analyzed to obtain time-accurate information about movements, actions, and poses, available for further analysis.
- Using information extracted from video recordings, both video and position data can be enriched with additional metadata.
- Using computational methods for sports field registration, position data can be extracted from videos.
- In the future, real-time approaches may help to evaluate live actions in individual and team sports.

4.1 Example Sport

Video recordings of competitions and training processes capture actions, poses and movements of athletes. They contain valuable information, e.g., body poses, body language (emotions, exhaustion, etc.) and biomechanical or tactical details, which cannot be represented by position data alone (see ► Chap. 6). Based on three examples from the domain of soccer, we illustrate how video recordings can be automatically analyzed with the help of AI-based approaches.

1. Video recordings of soccer matches contain a variety of atomic (fouls, goals, etc.) and complex (passes, shots, etc.) actions. In addition, details of the execution of actions are also visible, such as the part of the body with which a pass or shot is performed or the type of a pass played (flat pass, high pass). Methods for the automatic detection of actions, poses and position data can be used to

- enrich available information with further metadata. This allows analysts to search for scenes with specific actions in videos and large video collections, for example, in order to answer research questions from sports science efficiently.
2. Furthermore, the pose and movement patterns of athletes can be analyzed in more detail. For example, pose information such as the body and head orientation of a soccer player executing a pass, as well as possible pass receivers and defending players, can be taken into account to predict passing options. Methods for automatic object and motion tracking can be utilized to describe movement patterns as time series data.
 3. The importance of position data for various questions of game analysts and sports computer scientists is also described in ► Chap. 6. However, position data are usually not freely available and the creation of such position data requires special mobile devices or several cameras up to pre-installed camera systems in the stadium, which is typically only feasible in the professional domain. However, using a video recording from a single camera perspective and information about the playing field (field boundaries), it is possible to estimate position data automatically. In this way, analyses can also be realized for the amateur sector or for the analysis of training processes.

4.2 Background

On the one hand, position data help to objectively evaluate games relatively quickly by means of an abstract representation, but on the other hand, lots of details are naturally lost in the process. This includes information about actions (header, straddle, etc.), pose, head and gaze direction, as well as movement details. To capture such information, computer vision approaches can be applied to video data, usually based on deep learning models (see ► Chaps. 20 and 21). In recent years, several methods have been presented to detect actions in sports videos with exact time points (Biermann et al., 2021; Deliège et al., 2021; Giancola & Ghanem, 2021). Body pose estimation approaches (Cao et al., 2021; Kreiss et al., 2019) detect key points, e.g., shoulders, hips, knees, joints, etc. for the depicted subjects to represent the pose. These approaches are trained with videos from the domains of team and individual sports, among others, and have been successfully applied to different sports, for example, to evaluate possible pass options (Sangüesa et al., 2020) or to analyze penalty kicks (de Sousa Pinheiro et al., 2022).

Position data enable various other applications. Therefore, recently, more and more approaches have been proposed for sports field registration (Chen & Little, 2019; Sha et al., 2020; Theiner et al. 2023), which is important for the extraction of position data from videos. Methods for sports field registration transform the vis-

ible part of the pitch in the image or video frame into a 2D model of the sports field. In combination with approaches for object detection (Zhou et al., 2020), Theiner et al. (2022) have presented a first system for extracting position data from television and scouting feed recordings of soccer matches.

In addition to the previously mentioned research topics, there are more research fields around AI-based sports video analysis. These include, for example, the automatic generation of highlight videos (Decroos et al., 2017) and the tracking and (re)identification (even across shots taken from different cameras) of athletes or game equipment (Cioppa et al., 2022; Habel et al., 2022; Rematas et al., 2018). Furthermore, researchers have explored software tools (e.g., SportSense by Probst et al., 2018) and information visualizations (Fischer et al., 2019) for effective analysis processes using sports videos. Besides the aforementioned research approaches, various commercial tools, such as Skillcorner (► <https://skillcorner.com/>) and Stats Perform (► <https://www.statsperform.com>), also provide solutions for the analysis of sports videos.

4.3 Basics and Definition

Digital videos consist of a series of frames recorded at a frequency of usually 25–100 frames per second. For the human perception of smooth motion, at least 15 frames per second are necessary. Video data are therefore very storage-intensive. A video with a spatial resolution of 1920 * 1080 pixels with three color channels for the primary colors red, green and blue with eight bits each already requires about 178 MByte per second at a frame rate of 30 Hz without compression. For this reason, lossy compression methods are used that exploit the redundancy between rapidly successive images and can estimate movements from image to image. These compression methods can significantly reduce the amount of data required while largely preserving quality. For post-processing of videos, video editing software is necessary (editing software, editing program), which allows users, for example, to select individual video segments, to join them in a different order, to improve the image quality, or to insert or superimpose text information over the image content.

Videos of competitions and training sessions can be recorded in several camera settings with different characteristics. In SoccerNet (► <https://www.soccer-net.org/>), introduced by Deliège et al. (2021) which is one of the largest data collections in soccer, a distinction is made between 13 camera settings. The settings range from a main camera, which covers most of the field, to cameras for close-ups and goal cameras. Furthermore, there are also scouting feed recordings of soccer matches, which usually cover (almost) the entire sports field and are therefore particularly suitable for tactical analyses. Close-ups, on the other hand, depict certain motion sequences in greater detail and are therefore suitable for biomechanical analyses, for example.

Study Box

Theiner et al. (2022) presented a first system (Fig. 4.5.1) that combines state-of-the-art computer vision techniques to automatically estimate position data in television broadcasts and scouting feed recordings of soccer matches. For this purpose, the sports field and field markings in the video frame are segmented and compared to a reference database of synthetic images of the sports field with known camera parameters using a deep learning approach (Chen & Little, 2019). A homography matrix is determined based on the camera parameters of the most similar reference image. Subsequently, the homography matrix can be applied to transform the video frame into a 2D sports field model. This so-called sports field registration was further optimized

by Theiner and Ewerth (2023) by estimating camera parameters via an iterative optimization of the reprojection errors of geometric primitives (line segments of the sports field) to the 2D sports field model. Finally, the players are detected in the video images using a deep learning approach (Zhou et al., 2020) and transformed to the 2D sports field model through the homography matrix to estimate position data. The system has achieved very good results and provides an initial basis for performing various analyses based on position data (see Chap. 6), such as automatic classification of soccer formations (Müller-Budack et al., 2019), spatial control (Memmert et al., 2019), and other key performance indicators (KPIs).



Fig. 4.5.1 Processing pipeline for position data extraction from videos (Theiner et al., 2022)

4.4 Applications

► Example 1 (Action Detection)

AI-based approaches to automatically detect actions in videos can enrich video and position data with additional meta-information. The SoccerNet dataset from Delière et al. (2021) contains 500 videos annotated for 17 important actions in soccer, including goals, fouls, and shots on and off the goal. Current deep learning approaches (e.g.,

Giancola & Ghanem, 2021) already achieve promising results in spotting actions in videos. Using such approaches, it is possible to efficiently search for specific actions in videos and large video collections. For example, standard situations or situations that led to a goal could be analyzed in more detail for selected teams, matches, etc. On the other hand, starting from a selected scene, it is possible to find similar scenes in terms of number and type of actions in the same or in other videos. This allows game analysts to search for specific tactical patterns. An existing limitation of the SoccerNet dataset is that important actions such as passes as well as their attributes (e.g., whether flat or high pass) are not annotated. A corresponding taxonomy for a more complete coverage of actions in invasion sports has been presented by Biermann et al. (2021). With the help of annotated training data according to this taxonomy, current methods for action spotting (e.g., Giancola & Ghanem, 2021) can be finetuned and extended with these classes in the future. ◀

► Example 2 (Pose Information)

Video recordings of athletes contain important details about body pose and movements. Sangüesa et al. (2020) apply the deep learning model from OpenPose (Cao et al., 2021) to extract pose information from video data of soccer matches to determine the body orientation of the passer and possible pass receivers. The position data are enriched with body pose meta-information to predict the most likely pass options considering the positions of defending players. The method was evaluated based on 11 matches from the FC Barcelona with a total of 6038 passes. By combining body orientation information with position data, better results were obtained in predicting the pass recipient when compared to a reference method that used only position data. Thus, using a multimodal model resulted in better pass predictions, whereby body orientation was a main relevant feature for this improvement. ◀

► Example 3 (Player Detection and Tracking)

AI-based methods for (re-)identification allow us to track athletes across video frames and camera angles. Cioppa et al. (2022) published a dataset of 200 video sequences from 12 soccer matches, each of 30 seconds length, optimizing several state-of-the-art methods for multi-object tracking. By tracking players, movements can be extracted from video data in the form of trajectories, which can be used, for example, to analyze running patterns or movements. Especially when combined with methods for sports field registration (see study box), the movements can also be mapped to a 2D model of the sports field to estimate biomechanical performance indicators (e.g., running distances, running speeds) and perform further analyses based on position data (see ► Chap. 6). ◀

? Questions for the Students

1. What are the special characteristics of video data?
2. What are the advantages and disadvantages of video data over position data?
3. Which tasks can be solved by applying AI-based approaches in sports videos?

References

- Biermann, H., Theiner, J., Bassek, M., Raabe, D., Memmert, D., & Ewerth, R. (2021). A unified taxonomy and multimodal dataset for events in invasion games. *International Workshop on Multimedia Content Analysis in Sports co-located with the ACM Multimedia, MMSports@MM 2021, Virtual Event, 2021* (pp. 1–10). ACM.
- Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2021). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 172–186. IEEE.
- Chen, J., & Little, J. J. (2019). Sports Camera Calibration via Synthetic Data. *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, 2019* (pp. 2497–2504). IEEE.
- Cioppa, A., Giancola, S., Deliège, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., & Droogenbroeck, M. V. (2022). SoccerNet-Tracking: multiple object tracking dataset and benchmark in soccer videos. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, 2022* (pp. 3490–3501). IEEE/CVF.
- de Sousa Pinheiro, G., Jin, X., Da Costa, V. T., & Lames, M. (2022). Body pose estimation integrated with notational analysis: A new approach to analyze penalty kicks strategy in elite football. *Frontiers in Sports and Active Living*, 4.
- Decroos, T., Dzyuba, V., Haaren, J. V., & Davis, J. (2017). Predicting soccer highlights from spatio-temporal match event streams. *AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 2017* (pp. 1302–1308). AAAI Press.
- Deliège, A., Cioppa, A., Giancola, S., Seikavandi, M. J., Dueholm, J. V., Nasrollahi, K., Ghanem, B., Moeslund, T. B., & Droogenbroeck, M. V. (2021). SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, Virtual Event, 2021* (pp. 4508–4519). IEEE.
- Fischer, M. T., Keim, D. A., & Stein, M. (2019). Video-based analysis of soccer matches. *International Workshop on Multimedia Content Analysis in Sports co-located with the ACM Multimedia, MMSports@MM 2019, Nice, France, 2019* (pp. 1–9). ACM.
- Giancola, S., & Ghanem, B. (2021). Temporally-aware feature pooling for action spotting in soccer broadcasts. *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, Virtual Event, 2021* (pp. 4490–4499). IEEE.
- Habel, K., Deuser, F., & Oswald, N. (2022). CLIP-reIdent: Contrastive training for player re-identification. *International Workshop on Multimedia Content Analysis in Sports co-located with the ACM Multimedia, MMSports@MM 2022, Lisboa, Portugal, 2022* (pp. 129–135). ACM.
- Kreiss, S., Bertoni, L., & Alahi, A. (2019). PifPaf: composite fields for human pose estimation. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 2019* (pp. 11,977–11,986). IEEE.
- Memmert, D., Raabe, D., Schwab, S., & Rein, R. (2019). A tactical comparison of the 4-2-3-1 and 3-5-2 formation in soccer: A theory-oriented, experimental approach based on positional data in an 11 vs. 11 game set-up. *PLoS One*, 14, e0210191.
- Müller-Budack, E., Theiner, J., Rein, R., & Ewerth, R. (2019). “Does 4–4–2 exist?”—An analytics approach to understand and classify football team formations in single match situations. *International Workshop on Multimedia Content Analysis in Sports co-located with the ACM Multimedia, MMSports@MM 2019, Nice, France, 2019* (pp. 25–33). ACM.
- Probst, L., Kabary, I. A., Lobo, R., Rauschenbach, F., Schuldt, H., Seidenschwarz, P., & Rumo, M. (2018). SportSense: User interface for sketch-based spatio-temporal team sports video scene retrieval. *ACM Conference on Intelligent User Interfaces Workshops, ACM IUI Workshops 2018, Tokyo, Japan, March 11, 2018* (Vol. 2068). CEUR-WS.org.
- Rematas, K., Kemelmacher-Shlizerman, I., Curless, B., & Seitz, S. M. (2018). Soccer on your tabletop. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 2018* (pp. 4738–4747). IEEE.

- Sangüesa, A. A., Martín A., Fernández, J., Ballester, C., & Haro, G. (2020). Using Player's Body-Orientation to Model Pass Feasibility in Soccer. *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, 2020* (pp. 3875–3884). IEEE/CVF.
- Sha, L., Hobbs, J. A., Felsen, P., Wei, X., Lucey, P., & Ganguly, S. (2020). End-to-end camera calibration for broadcast videos. *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 2020* (pp. 13,624–13,633). IEEE.
- Theiner, J. & Ewerth, R. (2023). TVCalib: Camera Calibration for Sports Field Registration in Soccer. *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, 2023 (1166–1175)*. IEEE/CVF.
- Theiner, J., Gritz, W., Müller-Budack, E., Rein, R., Memmert, D., & Ewerth, R. (2022). Extraction of positional player data from broadcast soccer videos. *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, 2022* (pp. 1463–1473). IEEE/CVF.
- Zhou, X., Koltun, V., & Krähenbühl, P. (2020). Tracking objects as points. *European Conference on Computer Vision, ECCV 2020, Glasgow, UK, 2020* (pp. 474–490). Springer.



Event Data

Marc Garnica Caparrós

Contents

- 5.1 Example Sport – 36**
- 5.2 Background – 37**
- 5.3 Application – 38**
 - 5.3.1 Event Data to Extend Box Score Statistics – 38
 - 5.3.2 Event Data to Value in-Game Actions and Player Impact – 39
 - 5.3.3 Event Data to Understand Player Interactions – 39
- References – 40**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

5

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Event data is the time-ordered collection of all actions occurring in an invasion sports game such as soccer or basketball.
- Event data not only includes time and sport-specific attributes but also allocates all the events in the field and can contain custom attributes to enrich the data analysis.
- Invasion sports are a complex interaction of several players. Event data can help disseminate the sequences of a game and understand the outcome.
- Event data allows computing advanced statistics of a game that can contextualise the activity of certain players in certain situations.
- Several probabilistic models were presented in recent years aiming to analyse the chain of events that lead to a certain team objective, such as the Expected Goals metric.
- Not all goals are created equally, event data contains crucial information in understanding what, where and how it happens.
- A correct management and modelling of event data can contribute to a better understanding of team tactics and player contribution.

5.1 Example Sport

Not that long ago, soccer players were mainly evaluated only by their scoring ability. Player awards were then related to the players with higher offensive activity and efficiency. Prior to the explosion of interest in performance statistics and key indicators of players' contribution, the assist (i.e., the pass that enables another player to score) and the passing ability overall started receiving more attention. Nowadays and through the most prolific years of analytics in soccer and other sports, players

are evaluated on both offensive and defensive ends in high granularity. The short sequence between a goal and his preceding pass has been extended substantially, how was the attacking play started? Who moved the ball to the attacking zone where the assist was performed? The coaching staff no longer attributes player performance only to the last movements before a player scores, but to a larger and more meaningful sequence of events leading to it. This sequence of XY-located events in the pitch is the basis of Event Data, a time-based log of all actions occurring during the game. The analysis of event data has become a crucial aspect of any professional soccer team with the main applications sourcing from pattern mining, sequence analysis and association rules. In soccer, event data led to the emergence of advanced metrics such as the Expected Goals (xG) (Caley, 2015) and other extensions not only in soccer but also in sports like Basketball, where event data is also one of the data sources most used in daily operations, with expected metrics such as Expected Possession Value (Cervone et al., 2014).

5.2 Background

The emergence of highly fined granular data is one of the main motors of the big data analytics revolution that the sports industry experienced in the last 10 years. When watching or playing invasion sports (Hughes & Bartlett, 2002), sports with such common characteristics as soccer, there are several options to recreate the game through data. Box-score statistics, often called match sheet data, provide a very intuitive picture of the actions that occurred during the game, for example, the number of passes of a certain team in a basketball game (Oliver, 2004) or the number of shots in a soccer game. However, this data shows a discrete summary neglecting the interactions, time order and distribution of the game. The temporal information, i.e., when the events are happening and in which order, was added into the so-called *play-by-play* data, an ordered textual collection of all the actions performed by both teams during a game. This type of data source not only provided a more exact description of the game but also allowed for sequence-based analysis of events (Carling et al., 2008), moving the attention towards the chain of events rather than the appearance of an individual event.

Play-by-play data collection motivated the study of temporal interactions in invasion sports. In some cases, processing of this textual log could generate ad-hoc box scores with temporal criteria. For instance, the distribution of the events in time could give information about the sports structure and characteristics (Alberti et al., 2013). Despite the increased information present in this data source, the textual information of each action occurring in the game was often insufficient, limiting its application and studies. Thanks to the advances in data capture technologies and computer-vision systems (Gudmundsson & Horton, 2018), *play-by-play* data evolved into the so-called *event data*. Event data allows for a better understanding of the invasion sports game as it includes spatiotemporal information of all actions

occurring during the game as well as several context markers. The appearance of spatiotemporal event data in invasion sports is often related to another data source, tracking data or positional data (Bourbousson et al., 2010; Goes et al., 2020). Positional data is a highly granular data source collected by optical tracking systems or sensor-based technologies. This data source includes the locations of all players and the ball at a high frequency. Positional data includes more information than event data but it's often harder to analyse efficiently. Event data is commonly used for coaching, scouting or performance analysis purposes and has become a core component of any data-driven sports organizations, namely clubs and federations. Despite that most of the event data used is currently being collected by a mixture of manually annotated procedures and automatic systems, the advances in the automated notation of sports games are expanding this data source to all leagues and academies (Biermann et al., 2021).

Definition

Event data is defined as the time-ordered collection of all actions (events) that are occurring in an invasion sports game. The event information includes but is not limited to systematic information such as the timestamp when the event occurred, the key actor of the event (e.g., the player performing the pass), the team, the spatiotemporal features of the event (i.e., x and y coordinates of the event in the playing field) and the outcome of the event (e.g., whether the pass was accurate or not); and also sport-specific attributes. For instance, in soccer, events could be enriched with the part of the body used to perform the action (left foot, right foot, head), the type of event (e.g., diagonal passes, through passes, chipped passes), or the difficulty associated with the event (e.g., the number of defenders in front, the position of the goalkeeper when shooting, etc). Current advances in event data in soccer included on each event the location of all the players and the ball as a key context attribute (STATSBOMB, 2021). Event data is present in many invasion sports such as soccer, basketball, handball, hockey and rugby.

5.3 Application

5.3.1 Event Data to Extend Box Score Statistics

► Example 1

The raise of Women's soccer has been established in recent years as the guide for so many other sports organizations to boost and motivate equality in sports. From a performance point of view, a recent study compared the technical and tactical differences between

men's and women's soccer using event data (Garnica-Caparrós & Memmert, 2021). In order to extract the most detailed statistics to summarize a soccer game and enrich the comparison, the study made use of event data from two competitions (51 games of the 2016 UEFA Men's Europe Championship and 31 games of the 2017 UEFA Women's Europe Championship). 33 discrete features were created from over 100 K events disseminated by the period of the game and player position. A subjective comparison methodology was present by using machine learning interpretability tools. Overall, the study showcased pivotal factors that differentiate each gender's performance as well as patterns involving several indicators. ◀

5.3.2 Event Data to Value in-Game Actions and Player Impact

▶ Example 2

In order to extend the approaches led by the Expected Goals (xG) metric, the full potential of event data was used with the goal to measure the impact of every action in a soccer game (Decroos et al., 2020). The VAEP framework tries to assign a contribution to every single event by measuring the probabilities of scoring and conceding a goal preceding and proceeding with the event. In doing so, this approach aimed at improving existing methods that only rely on rare events, such as shots, to evaluate a player or team's performance. Overall, the VAEP framework can be used to quantify a player's or team's offensive and defensive contributions. ◀

5.3.3 Event Data to Understand Player Interactions

▶ Example 3

The information that event data comprises enables to analyse team's behaviour as a complex system of interactions. (Duch et al., 2010) highlight the power of network analysis to understand the interaction between players in a team. Passing networks were presented as a visual definition of teamwork and quantified the contribution of individuals and team performance. A passing network can be built from simple Event Data sources, the nodes of the network represent the players of a team and the edges represent their connection during the game (i.e., passes). Weighted edges provide an overview of the most common interactions and pivotal players on team tactics. If available, nodes can also be allocated in the average position of the players in the field, providing an overall allocation of the team in the field and an XY-based overview of their tactics. ◀

Study Box

Current research on the use of event data sources to understand invasion sports is driven by the probabilistic models extending the work of the xG metric and the VAEP framework. However, its predominant use in sports organizations forms a need for a more democratic, interpretable and customizable application of sequence-based analysis of event data. A recent study (Kröckel & Bodendorf, 2020) proposes a generalizable framework for players' contribution analysis, team tactics and sequence anal-

ysis in soccer that could be extrapolated to the rest of invasion sports. Process Mining as a tool could serve as an entry point to unfold the potential of detailed event data sources beyond computing advanced box score statistics. Process-based management could also allow the reproducibility of sequences in experimental approaches through simulation. Oversampling and artificial modelling of event sequences could expand the existing knowledge in expected metrics algorithms and refine their interpretation.

5

? Learning Control Questions

1. What is the main difference between *play-by-play* and event data?
2. How does event data improve box score or match sheet information?
3. How is the VAEP framework able to measure the contribution of each action in the game?

References

- Alberti, G., Iaia, F. M., Arcelli, E., Cavaggioni, L., & Rampinini, E. (2013). Goal scoring patterns in major European soccer leagues. *Sport Sciences for Health*, 9, 151–153. <https://doi.org/10.1007/s11332-013-0154-9>
- Biermann, H., Theiner, J., Bassek, M., Raabe, D., Memmert, D., & Ewerth, R. (2021). A unified taxonomy and multimodal dataset for events in invasion games. <https://doi.org/10.48550/ARXIV.2108.11149>.
- Bourbousson, J., Sève, C., & McGarry, T. (2010). Space–time coordination dynamics in basketball: Part 2. The interaction between the two teams. *Journal of Sports Sciences*, 28, 349–358. <https://doi.org/10.1080/02640410903503640>
- Caley, M. (2015). EPL projections and expected goals method: Spurs are good! *EPL projections and expected goals method: Spurs are good!* Retrieved from <https://cartilagefreecaptain.sbnation.com/2015/10/19/9295905/premier-league-projections-and-new-expected-goals>
- Carling, C., Bloomfield, J., Nelsen, L., & Reilly, T. (2008). The role of motion analysis in elite soccer. *Sports Medicine*, 38, 839–862. <https://doi.org/10.2165/00007256-200838100-00004>
- Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2014). *A multiresolution stochastic process model for predicting basketball possession outcomes*. <https://doi.org/10.48550/ARXIV.1408.0777>
- Decroos, T., Bransen, L., Haaren, J. V., & Davis, J. (2020). VAEP: An objective approach to valuing on-the-ball actions in soccer (extended abstract). *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2020/648>.
- Duch, J., Waitzman, J. S., & Amaral, L. A. (2010). Quantifying the Performance of Individual Players in a Team Activity. (E. Scalas, Ed.). *PLoS One*, 5, e10937. <https://doi.org/10.1371/journal.pone.0010937>

- Garnica-Caparrós, M., & Memmert, D. (2021). Understanding gender differences in professional European football through machine learning interpretability and match actions data. *Scientific Reports*, *11*. <https://doi.org/10.1038/s41598-021-90,264-w>
- Goes, F. R., Meerhoff, L. A., Bueno, M. J., Rodrigues, D. M., Moura, F. A., Brink, M. S., et al. (2020). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, *21*, 481–496. <https://doi.org/10.1080/17461391.2020.1747552>
- Gudmundsson, J., & Horton, M. (2018). Spatio-temporal analysis of team sports. *ACM Computing Surveys*, *50*, 1–34. <https://doi.org/10.1145/3054132>
- Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of Sports Sciences*, *20*, 739–754. <https://doi.org/10.1080/026404102320675602>
- Kröckel, P., & Bodendorf, F. (2020). Process mining of football event data: A novel approach for tactical insights into the game. *Frontiers in Artificial Intelligence*, *3*. <https://doi.org/10.3389/frai.2020.00047>
- Oliver, D. (2004). *Basketball on Paper*. Potomac Books Inc..
- STATSBOMB. (2021). *STATSBOMB*. Retrieved from <http://www.statsbomb.com/>



Position Data

Daniel Memmert

Contents

- 6.1 Example Sport – 44
- 6.2 Background – 45
- 6.3 Applications – 46
- References – 47

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

6

Key Messages

- Position data describe the positions/movements of athletes and game equipment in terms of X–Y coordinates
- Position data reflect the complex reality, are reliable, can be evaluated objectively and extremely quickly
- Performance-relevant parameters from training and competition can be analyzed on the basis of position data
- Experimental approaches can help develop and test theories in various areas of sports science and sports informatics in the future

6.1 Example Sport

Three examples from the sports game soccer will be used to illustrate the questions that game analysts—but also sports computer scientists—have in a wide variety of team sports games that can be answered on the basis of position data (Garnica-Caparrros & Memmert, 2021; Rein et al., 2017b). First, it is important to know which spaces on the soccer field are occupied by which games and when. In addition to such so-called space control values, it is also of interest how control shares shift with one's passing game: How big is the space gain in the game setup or in interface passes in front of the opponent's goal? Secondly, when it comes to a team's switching behavior, it is important to know how quickly and where a team's players run at the opponent after losing the ball. In this way, not only the success rate in pressing can be quantified, but also the aggressiveness with which a team switches after losing the ball. Thirdly, you can determine how many opponents can still defend a player with the ball—both before and after he has played a pass. The two values already provide information about how many players a team puts

behind the ball when defending (rest defense), but the difference is just as interesting: it shows how many opponents ultimately overplay a pass and thus take it out of play. This value proves to be a good method for evaluating vertical passes and can be refined as desired by, for example, integrating the pressure that the opponents apply to the passer and receiver at the time of delivery and reception in addition to the overplayed opponents.

6.2 Background

To capture the performance parameters presented above, it would take many hours to evaluate video recordings. On the basis of so-called position data, however, this is possible in seconds (Memmert & Raabe, 2019). The precise recording of the positions of each actor and possibly other materials thus enables significantly more complex analyses with new types of performance indicators today. Meanwhile, position data are collected in various sports. While in (beach) volleyball (Link, 2014), field hockey (Stöckl & Morgan, 2013), handball (Hassan et al., 2017), tennis (Kovalchik & Reid, 2018; van Meurs et al., 2021), badminton (Rojas-Valverde et al., 2020) or basketball (Kempe et al., 2015) this development is only at the beginning, in soccer position data (Theiner et al., 2022) are already generated by default (Biermann et al., 2023; cf. ► Chap. 5). There are several performance measures, known as key performance indicators (KPIs), that have been used for analysis to date (see Low et al., 2019; Memmert et al., 2017).

With the help of process, longitudinal and cross-sectional analyses, it is also possible to investigate dynamic relationships on the basis of position data, for example, by taking greater account of situational and thus context-specific references and interindividual differences (cf. Rein & Memmert, 2016). For this purpose, training and competition data can also be more strongly linked in the future. In order to interpret positional data, sound theories or models are mandatory (Memmert et al., 2019; Rein et al., 2017a). Due to the ever-increasing importance of positional data in sports informatics, experimental approaches that establish theories and test them empirically are currently becoming visible (Low et al., 2022a, 2022b; Memmert et al., 2019).

Definition

Position data describe the positions and movements of athletes and playing equipment in the form of X–Y coordinates. In sports games, they consist of the positions of all players and the ball in the form of X–Y coordinates (for the ball, sometimes also Z components) (Memmert & Raabe, 2019). They are recorded either by special camera systems in the stadium or by mobile devices worn by the players under their clothes (Memmert, D. 2021. Match Analysis. Abingdon: Routledge).

Study Box

The field experiment by Memmert et al. (2019) is the first to investigate the effects of different playing systems (here 4-2-3-1 vs. 3-5-2) on tactical KPIs using positional data in an 11 vs. 11 soccer match setup. The KPIs were measured using dynamic KPIs such as “Effective Playing Area,” “Length-to-Width Ratio,” space control, and passing efficiency under pressure.

Within the experimental positional data analysis paradigm, both team formations showed no differences in effective

playing space or space control. Consistent with the hypothesis, a 3-5-2 playing system with 5 levels (3-1-2-2) outperformed the 4-2-3-1 playing system with 4 levels for the “length-width quotient” and passing efficiency under pressure, because the former had one more (player) level in the mid-field. The experimental paradigm for positional data analysis represents a useful approach to advance the development and validation of theory-based models in sports game performance analysis.

6

6.3 Applications**► Example 1**

In a Big Data field study, a total of 50 matches of the men’s soccer Bundesliga from the 2014/15 season (2 teams, 2 half-times, 200 data sets) were automatically evaluated and validated based on positional data with different KPIs (Memmert et al., 2016, 2017). The focus was on the self-developed analysis tool SOCCER (Perl et al., 2013), which combines conventional data analysis, dynamic state-event modeling, and artificial neural networks (cf. ► Chap. 20). The winning teams convince with significantly higher space control shares as well as space gains in their own build-up of play and also outplay more opponents here on average. They also win a lot of space in front of the opponent’s goal in attack. Over the course of the entire season, the teams in the top and bottom third of the table (according to the final table) were also compared with each other. And here, too, space control proved to be a major difference between top clubs and relegation candidates. Regardless of whether it was the build-up to the game or attacking play: in almost all areas, there was a significant difference in spatial dominance in the critical zones of the pitch—in favor of the teams from the top third. If we compare the winning and losing teams, we see that the winning team overplays more opponents in the game setup during the 90 min. They also face fewer opponents on average in possession compared to the losers—even on vertical passes in the attacking area. Nevertheless, the losing teams made more effort in the transition game. ◀

► Example 2

In order to mask different physiological and anatomical characteristics between women and men and to avoid gender bias in the assessment of soccer matches at the highest level of play, tactical performance of both genders was assessed based on positional data (Garnica-Caparros & Memmert, 2021). Artificial neural networks (s. ► Chap. 20) were used as objective KPIs, among others. The analysis of pass pressure efficiency measures,

different pressing indices as well as different space control parameters reveals that women and men show comparable values in all tactical variables. In summary, it was shown that, in contrast to previous video-based analyses, no significant differences in soccer-specific tactical performance between women and men in high-performance soccer are detected when using “blinded” positional data, where no inferences can be made about gender, and objective KPIs. The results can be used to provide objective conclusions about the training of players, contribute to the further development and professionalization of women’s soccer in the area of tactics, and help to promote the public perception and attractiveness of women’s soccer on the basis of objective evaluation criteria. ◀

► Example 3

Including new context information during match phases, Klemp et al. (2022) investigated the relationship between running performance and goal scoring in professional soccer. In a sample of 302 matches of the first Bundesliga, the first goal was modeled as a function of running performance, based on positional data, and team strength of the teams using logistic regression. The best model showed a median accuracy of 77%, reflecting a strong relationship between running distance and the probability of scoring the first goal. This relationship was strongest for total running distance compared to sprinting distance or running distance with own ball possession. The authors propose two different possible mechanisms to explain the relationship between running performance and scoring success found in the present study. On the one hand, better fulfillment of the players’ tactical goals could be responsible for this, on the other hand, the increasing fatigue of the opposing players may also play a role. ◀

► Example 4

Guerrero-Calderón et al. (2021) analyzed the physical performance of professional soccer players during training, taking into account the contextual factors of match location, season duration, and opponent quality, in order to build predictive models for the performance delivered during training sessions. Training data were generated from 30 professional soccer players of the Spanish La Liga based on positional data ($N = 1365$ performances). During the training weeks prior to home matches, reduced effort was shown in terms of various strength, speed, and endurance parameters. The quality of the opponent also affected the training load. The proposed predictive model represents an innovative approach to quantify training load in professional soccer considering novel contextual factors. ◀

? Questions for the Students

1. What is position data?
2. Give two concrete examples of how it can be used in sports.

References

-
- Biermann, H., Komitova, R., Raabe, D., Müller-Budack, E., Ewerth, R., & Memmert, D. (2023). Synchronization of passes in event and spatiotemporal soccer data. *Scientific Reports*, 13(1), 15878.

- Garnica-Caparros, M., & Memmert, D. (2021). Understanding gender differences in professional European football through Machine Learning interpretability and match actions data. *Scientific Reports*, *11*(1), 1–14.
- Guerrero-Calderón, B., Klemp, M., Morcillo, J. A., & Memmert, D. (2021). How does the workload applied during the training week and the contextual factors affect the physical responses of professional soccer players in the match? *International Journal of Sports Science & Coaching*, *16*, 994–1003.
- Hassan, A., Schrapf, N., & Tilp, M. (2017). The prediction of action positions in team handball by non-linear hybrid neural networks. *International Journal of Performance Analysis in Sport*, *17*, 293–302.
- Kempe, M., Grunz, A., & Memmert, D. (2015). Detecting tactical patterns in basketball: Comparison of merge self-organising maps and dynamic controlled neural networks. *European Journal of Sport Science*, *15*, 249–255.
- Klemp, M., Memmert, D., & Rein, R. (2022). The influence of running performance on scoring the first goal in a soccer match. *International Journal of Sports Science & Coaching*, *17*, 558–567.
- Kovalchik, S., & Reid, M. (2018). A shot taxonomy in the era of tracking data in professional tennis. *Journal of Sports Sciences*, *36*, 2096–2104.
- Link, D. (2014). A toolset for beach volleyball game analysis based on object tracking. *International Journal of Computer Science in Sport*, *13*(1), 24–35.
- Low, B., Coutinho, D., Gonçalves, B., Rein, R., Memmert, D., & Sampaio, J. (2019). A systematic review of collective tactical behaviours in football using positional data. *Sports Medicine*, *50*, 343–385.
- Low, B., Schwab, S., Rein, R., & Memmert, D. (2022a). Defending in 4-4-2 or 5-3-2 formation? Small differences in footballers' collective tactical behaviours. *Journal of Sports Sciences*, *40*, 1–13.
- Low, B., Schwab, S., Rein, R., & Memmert, D. (2022b). The porous high-press? An experimental approach investigating tactical behaviours from two pressing strategies in football. *Journal of Sports Sciences*, *39*, 2199–2210.
- Memmert, D., Lemmink, K., & Sampaio, J. (2017). Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine*, *47*, 1–10.
- Memmert, D., & Raabe, D. (2019). *Revolution im Profifußball. Mit Big Data zur Spielanalyse 4.0 (2. aktualisierte und erweiterte Auflage)*. Springer.
- Memmert, D., Raabe, D., Knyazev, A., Franzen, A., Zekas, L., Rein, R., Perl, J., & Weber, H. (2016). Big Data im Profi-Fußball—Analyse von Positionsdaten der Fußball-Bundesliga mit neuen innovativen Key Performance Indikatoren. *Leistungssport*, *46*, 21–26.
- Memmert, D., Raabe, D., Schwab, S., & Rein, R. (2019). A tactical comparison of the 4-2-3-1 and 3-5-2 formation in soccer: A theory-oriented, experimental approach based on positional data in an 11 vs. 11 game set-up. *PLoS One*, *14*, e0210191.
- Perl, J., Grunz, A., & Memmert, D. (2013). Tactics analysis in soccer- an advanced approach. *International Journal of Computer Science in Sport*, *12*, 33–44.
- Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *Springerplus*, *5*, 1–13.
- Rein, R., Perl, R., & Memmert, D. (2017a). Maybe a tad early for a grand unified theory: Commentary on “towards a grand unified theory of sports performance” by Paul S. Glazier. *Human Movement Science*, *56*, 173–175.
- Rein, R., Raabe, D., & Memmert, D. (2017b). “Which pass is better?” Novel approaches to assess passing effectiveness in elite soccer. *Human Movement Science*, *55*, 172–181.
- Rojas-Valverde, D., Gómez-Carmona, C. D., Fernández-Fernández, J., García-López, J., García-Tormo, V., Cabello-Manrique, D., & Pino-Ortega, J. (2020). Identification of games and sex-related activity profile in junior international badminton. *International Journal of Performance Analysis in Sport*, *20*, 323–338.
- Stöckl, M., & Morgan, S. (2013). Visualization and analysis of spatial characteristics of attacks in field hockey. *International Journal of Performance Analysis in Sport*, *13*, 160–178.
- Theiner, J., Gritz, W., Müller-Budack, E., Rein, R., Memmert, D., & Ewerth, R. (2022). Extraction of positional player data from broadcast soccer videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. Waikoloa, HI, USA. pp. 1463–1473
- van Meurs, E., Buszard, T., Kovalchik, S., Farrow, D., & Reid, M. (2021). Interpersonal coordination in tennis: Assessing the positional advantage index with Australian Open Hawkeye data. *International Journal of Performance Analysis in Sport*, *21*, 22–32.



Online Data

Christoph Breuer

Contents

- 7.1 Example Sport – 50
- 7.2 Background – 51
- 7.3 Application – 52
- References – 54

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

7

Key Messages

- Through digitalization, more and more sports data and other data relevant to sports science are available as online data.
- Web scraping or web crawling is a method to make this data systematically available for sports science research
- The use of online data leads to significant knowledge in sports science.

7.1 Example Sport

Digital sports data holds an enormous potential with a huge variety of applications, ranging from (1) enriched sports media content via (2) economical price strategies up to (3) individual training data. In order to give some examples, sports media witnessed an increase in online sports reporting and countless second-screen offers. Additionally, more and more TV broadcasters and streaming providers offer a tactical feed parallel to their classic sports broadcast, whereas sports leagues and betting providers give free access to performance statistics. Digitalization enables sporting goods manufacturers to apply dynamic price strategies via their online selling points while countless training, health, and running apps provide individual performance measurements.

Accordingly, this enables a myriad of possibilities for data-based sports research. Pioneering work has been done by Kemper and Breuer (2016a), who quantify dynamic pricing potentials by applying second sales ticket prices of FC Bayern München home tickets. As such, the authors showed that FC Bayern München would benefit from introducing a dynamic pricing system widely used in North American major league sports.

7.2 Background

Whereas online data vary significantly in its content, technically, the data is commonly available via online platforms. In order to use data for academic research, researchers may either copy the data manually (manual scraping) or via web scraping, where data extraction, copying, storing, and recycling are performed automatically. In this regard, the term web crawling is also used in academic literature.

With secondary data analysis, copyright and similar legal issues must be considered. Thus, data's copyright status should always be verified before researchers start the scraping. As a rule of thumb, all information accessed with a username and password is considered private and should not be analyzed (Bradley & James, 2019).

Initially, a basic knowledge of programming language skills (e.g., Python) was necessary to apply web scraping. Modern web scraping tools such as Octoparse, Parsehub, Scraper API, or Scarpe Simple allow users to create web scrapers, even without programming knowledge. Also, the latest updates in statistic programs such as GNU R provide packages to perform web scraping in the program's interface. Still, the python-based automated information extraction trumps any web scraping tools. Javascript can make web scraping more difficult. But there are software solutions for this too, such as PhantomJS.

Naturally, web scraping is not always welcomed. Some website domains block web scraping since massive web scraping affects website performance significantly. However, an increasing trend of legally usable online data is recognized.

Technically, the web scraping process follows four steps: First, the algorithm obtains the website's URL. Second, the web scraper retrieves and stores the website's HTML code. Third, the retrieved HTML code is now used to identify the interesting elements stored in a table or database. Finally, the command is adapted in accordance with the interesting elements.

Unfortunately, providing data in a structured, machine-readable, and secured Application Programming Interface format (API) is not common. At least in sports online data. Such API represents a technically and legally more secure procedure and will gain more importance in the future. A role model in this field is the National Basketball Association (NBA) which provides API interfaces to support data usage on its website ► nba.com. As such, statistic programs such as the previously mentioned statistic program R offer user-written applications to analyze data without web scraping detours (e.g., nbastatR Bresler, 2021). Similarly, numerous API applications are evident in fantasy sports.

Definition

Generally, online data are text data published on the internet covering a huge content variety. Using web scraping, such online data can be automatically extracted, copied, stored, and used. In this context, API interfaces are a legally more secure alternative but less available in sports.

Study Box

In 2009, the San Francisco Giants from the American Major League Baseball were the first sports club to introduce a dynamic pricing system. Since then, many North American professional sports teams have followed this example. However, no German sports club has yet introduced a dynamic pricing system. With the pioneering work done by Kemper and Breuer (2016b), a solid discussion foundation is given. In the research context, the authors collected secondary ticket prices from eBay's online auction website during the second half of the 2013/14 Bundesliga season. Excluding VIP seats and categorized customer groups (Students, seniors, and disabled persons) yielded in total 6510 eBay auctions in which 11,637 tickets were sold. Since many auctions sold multiple tickets

at once, the total price paid was divided by the number of tickets sold. Matching these eBay auctions with a variety of additional data (date and time from the official Bundesliga homepage, match results and table positions from ► ergebnisselive.com, derby information from ► derbysieg.com, spectator figures from fussball-daten.sport.de, population and capita data from MB-Research (2013), weather from dwd.de), produced an extensive data set of 6510 auctions, analyzed using a two-stage least squared regression. The results showed that tickets resold on the secondary market almost doubled the initial price of the tickets. Sports managers may further elaborate their current variable ticket pricing strategy to implement a more sophisticated dynamic pricing approach.

7

7.3 Application**► Example 1**

Traditionally, empirical studies on dynamic pricing in sports suffered from a limited number of observations. Applying online data and web scraping enabled Kemper and Breuer (2016b) to collect ticket prices daily over the entire sales period of a football club. As such, daily ticket prices of the English football club Derby County during the 2013/2014 season were listed. Analyzing ticket prices from the earliest possible purchase retrieved 5862 price points for 11 home games, considered for adults, seniors and U18 age groups. Obtaining prices directly from Derby County's official website (► <http://www.tickets.wearederby.com/match-tickets/buy-tickets/#>) and the number of sold tickets from ► worldfootball.net (► <http://www.worldfootball.net>) enabled a hedonic price regression which revealed a significant time effect on dynamic ticket prices. In more detail, the study found a monotone price-time relationship, which differs from pricing models in the aviation or hotel industry. However, sports managers may apply these findings to devise a more sophisticated pricing concept. ◀

► Example 2

In a sponsorship-related study, Breuer et al. (2021) quantified a live match's impact mechanism and moderator effect on TV viewers and the associated sponsors' benefits. Therefore, the spectators' physiological (gaze hits on advertising boards, electrodermal activity) and psychological (facial expressions) data were collected. Supplemented by live-betting odds to indicate the uncertainty of outcome, the authors showed for the first time how TV spectators' emotions vary along the game and how the advertiser's message perception is affected. ◀

► Example 3

In a similar study, Herold et al. (2021) examined the effect of so-called ghost games on TV spectators and sponsors (the COVID-19 pandemic required the Bundesliga to play football matches without spectators in the stadium). In this context, Herold et al. (2021) not only used web-scraped live-betting odds to control for the game's course and its uncertainty of outcome. The authors also obtain tension indicating match characteristics such as goals or associated match time. Web scraping technics were used to record the accessible bookmakers' betting odds immediately after being published online. Matching the betting odds with the participants' physiological data showed that ghost matches result in lower utility for TV spectators, albeit in already decided matches. Vice versa, ghost matches increase sponsors' benefit since spectators often glance at advertising media instead of stadium spectators' reactions. ◀

► Example 4

Steinfeldt et al. (2022) investigated whether spectators affect the score difference between the teams. The authors analyzed $n = 12,500$ NBA games from 11 seasons from 2010/11 to 2020/21. COVID-19 spectator restrictions granted the floor investigating a potential spectators' effect. Using the R package `nbastatR` (Bresler, 2021), data on a game level and regular season data were collected. The data set includes information of match results, location, teams' records, and a range of more sophisticated basketball statistics matched with spectator data scraped from ► [basketball-reference.com](https://www.basketball-reference.com) and betting odds from ► [oddsportal.com](https://www.oddsportal.com). The authors could demonstrate that games played with limited spectators were more likely to be won by margins of 15, 20, or 25 points than unrestricted crowds. Given that the effect was most severe for games played on a weaker team's home court, Steinfeldt et al. (2022) conclude that predominately weaker teams suffer from limited crowd support. ◀

? Questions for the Students

1. What are online data?
2. What should be considered in web scraping?

References

- Bradley, A., & James, R. J. E. (2019). Web scraping using R. *Advances in Methods and Practices in Psychological Science*, 2(3), 264–270.
- Bresler, A. (2021). nbastatR: R's interface to NBA data. R package version 0.1.1505. Retrieved November 24, 2021, from <https://github.com/abresler/nbastatR>
- Breuer, C., Rumpf, C., & Boronczyk, F. (2021). Sponsor message processing in live broadcasts—A pilot study on the role of game outcome uncertainty and emotions. *Psychology & Marketing*, 38(5), 896–907.
- Herold, E., Boronczyk, F., & Breuer, C. (2021). Professional clubs as platforms in multi-sided markets: The role of spectators and atmosphere in live football. *Sustainability*, 13, 2312.
- Kemper, C., & Breuer, C. (2016a). Dynamic ticket pricing and the impact of time—An analysis of price paths of the English soccer club Derby County. *European Sport Management Quarterly*, 16(2), 233–253.
- Kemper, C., & Breuer, C. (2016b). How efficient is dynamic pricing for sport events? Designing a dynamic pricing model for Bayern Munich. *International Journal of Sport Finance*, 11(1), 4–25.
- Steinfeldt, H., Dallmeyer, S., & Breuer, C. (2022). The silence of the fans—The impact of restricted crowds in the margin of victory in the NBA. *International Journal of Sport Finance*, 17, 165–177.



Modeling

Contents

- Chapter 8** **Modeling – 57**
Jürgen Perl and Daniel Memmert
- Chapter 9** **Predictive Models – 65**
Fabian Wunderlich
- Chapter 10** **Physiological Modeling – 73**
Manuel Bassek



Modeling

Jürgen Perl and Daniel Memmert

Contents

- 8.1 Example Sport – 58**
- 8.2 Background – 60**
- 8.3 Application – 62**
- References – 63**

Jürgen Perl was deceased at the time of publication.

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

8

Key Messages

The idea of modeling in sport is to map complex dynamic systems to their essential structures, data, and interactions in order to perform descriptive, prognostic, or planning analyses and calculations.

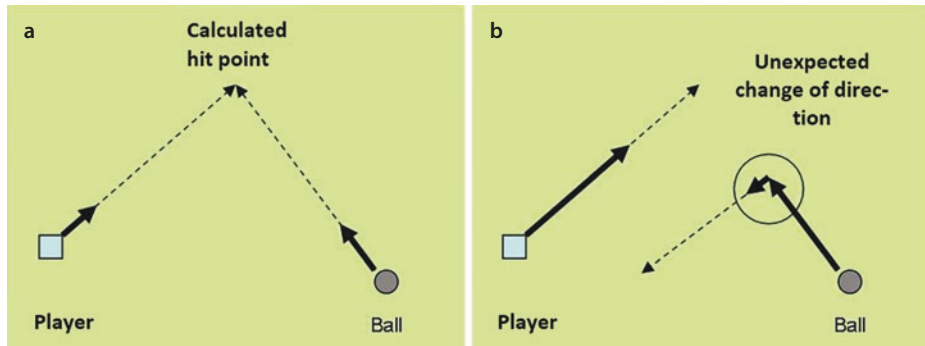
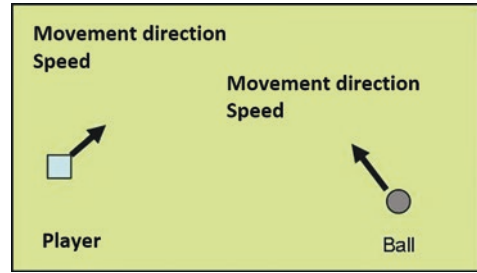
The four essential steps for developing and applying a model are:

- The *reduction* of the real system to its essential components and dynamics.
- The *mapping* of the problem situation to (informatically) manageable objects (numbers, functions, graphs, etc.) taking into account a corresponding back-mapping of the obtained results.
- The *analysis/calculation* is a process of question-oriented information extraction—or in short: the transformation of the problem or input data into result or output data.
- The *visualization* of the results on a system-oriented level (e.g. graphical representation).

8.1 Example Sport

Recognition of technical-tactical strengths and weaknesses or the development of strategic concepts based on captured position and action data is essential in soccer (Memmert, 2021). In the following, the case of model-based analysis of player-ball interaction is presented as an introductory example: "Could the player still have reached the ball?" For this purpose, ■ Fig. 8.1 shows a representation or visualization scheme in which the two objects selected in the reduction, player, and ball, are shown with their relevant data for the model (positions of the player and the ball

■ **Fig. 8.1** Visualization of the system reduction



■ **Fig. 8.2** (a) Result of the model calculation. (b) Comparison with reality

with their directions of movement at the beginning of the process as well as velocity values for player and ball from the video data).

The task of the model would now be to answer the input question of reachability from these data by appropriate calculation. The result of this model calculation (or simulation) as an answer to the reachability question would be “yes” or “no” in the simplest case. In both cases, however, the answer remains unsatisfactory because it does not convey where ball reachability occurs or why it does not occur. Thus, as the fourth step of modeling, adequate visualization of the results is necessary.

The calculated path graph shows in ■ Fig. 8.2a that the player theoretically could have reached the ball; but not practically: As the video recordings show, the ball was kicked away by an opponent before the calculated contact time (cf. ■ Fig. 8.2b). But this opponent was not part of the model, i.e. the model had reduced reality too much! And this brings us back to the first and decisive aspect of modeling—the reduction: The “reduction of the real system ...” mentioned above under (1) is necessary to be able to calculate a result at all and with reasonable effort. But it must not be too narrow, in order not to leave out essential objects and dynamics, which influence the result.

8.2 Background

The essential aspect of reduction for the mode of action and usability of a model for the soccer example can be seen in [Fig. 8.2b](#) (Perl & Memmert, 2019): The ball has undergone an abrupt change of motion in its course, which cannot be explained from the model, but is immediately understandable for the observer: the intervention of an opposing player. This opposing player was not part of the model because of too strong reduction, and therefore its possible effect on the motion dynamics to be modelled could not be recognized and calculated.

[Figure 8.3a](#) shows the typical modeling of such a player-opponent situation: Assuming that both players move with the same speed, the dividing line between the blue and the yellow area shows all points reached by the blue and the yellow player simultaneously. To all points of the blue area, the blue player reaches faster, to all points of the yellow area his yellow opponent does. These areas of faster reachability are also called the player's Voronoi cell after its "discoverer". The analysis of the reachability of the ball thus becomes more precise to the question (e. g., Rein et al., 2017): "How does the ball pass through the Voronoi cells of the two players?"

[Figure 8.3b](#) shows even if the blue player had moved in the optimal direction, he would not have had a chance to prevent the yellow opponent's action—he could not reach the ball before his opponent. Game analyses based on Voronoi cells are now standard in soccer and are used to analyze the effectiveness of tactical formations in terms of space control (Memmert & Raabe, 2018; Perl & Memmert, 2015) ([Fig. 8.4](#)).

Having thus shown that less reduction can also improve the accuracy of modeling, the question arises: should even more aspects of reality, such as speed differences and changes or changes in movement directions, be incorporated into the model? This question, which is central in modeling, cannot be answered with a

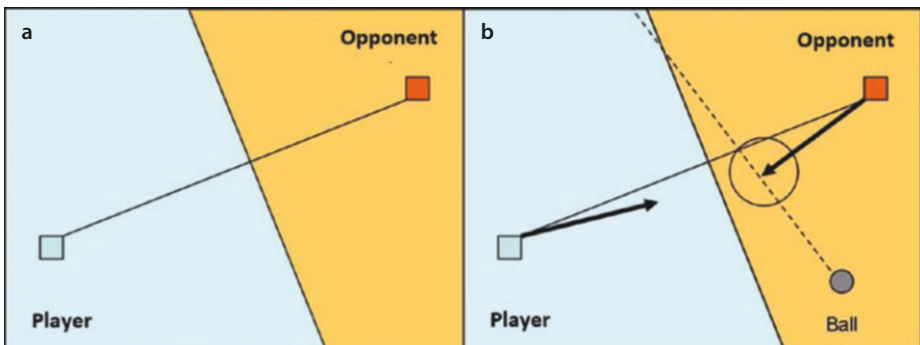


Fig. 8.3 (a) Players' Voronoi cells. (b) Reachability analysis

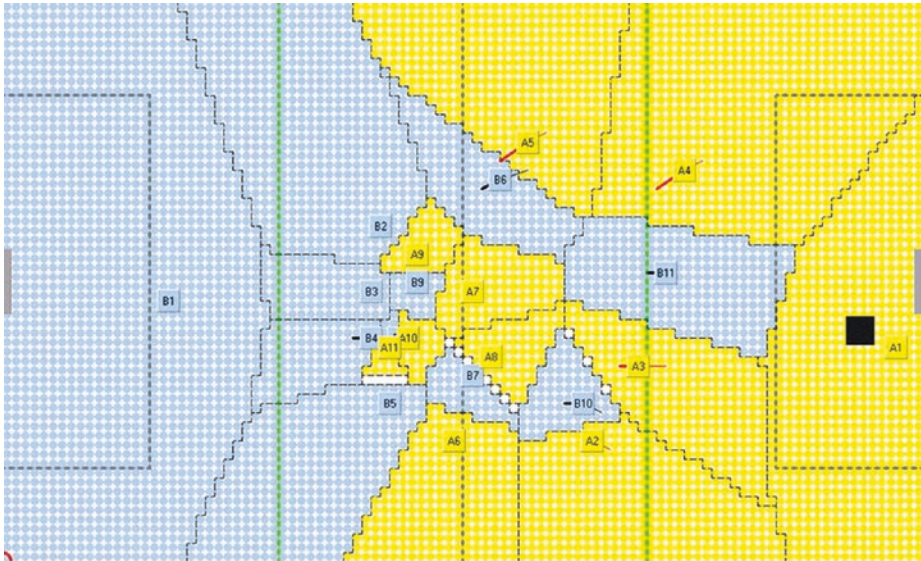


Fig. 8.4 Soccer field with the Voronoi cells of the players of “A-yellow” and “B-blue” (Memmert & Raabe, 2018)

blanket yes or no. The answer depends in each case on the available data, the still justifiable effort, and the expected benefit of modeling and calculation. For example, the aforementioned additions provide the possibility of a technical visualization of the game event parallel to the video presentation—but only if the data are available in sufficient scope and precision. Otherwise, the modeling visualizes the data deficits rather than the gameplay. **Résumé:** The central art of modeling is an adequate reduction that preserves essential dynamics without getting lost in gimmicks (Perl, 2015).

Definition

The model is an abstract representation of a system. It is used to diagnose the system state and predict the system behavior (Perl & Uthmann, 1997).

The 4 essential steps of modeling are (soccer example in parentheses):

- System reduction (capturing and representing the player-ball situation)
- Problem mapping (setting of position and velocity data)
- Analysis/calculation (calculation of the running paths and, if necessary, the intersection)
- Result visualization (representation of the player-ball situation as a graph)

Study Box

The goal of key performance indicators (KPI; Memmert et al., 2017, Low et al., 2019) is to map complex system behavior to single values in order to scale, score, and rank systems or system components. However, very often this mapping only reduces important information about tactical behavior or game dynamics without replacing it with more meaningful information. Perl and Memmert (2017) used a two-step approach to bridge the gap between complex dynamics and numerical metrics in offensive play in soccer. First, they developed a model that visualizes offensive action in a process-oriented manner by using KPIs to represent offensive performance. Second, this model has been organized in terms of time intervals, allowing effectiveness to be measured both for an entire half and for intervals of arbitrary length. In doing so, Perl and Memmert (2017) have shown that the attack efficiency profile is a dynamic indicator of a team's match success. In **Fig. 8.5**, red profiles show how the attack efficiency values of “A-yellow” and “B-blue” for the correlation interval length $IL = 300$ sec evolve over halftime. The efficiency values (OS A, OS B) for the second $I0 = 1721$ are plotted in the gray box. In the graph, the green profiles show the respective space control proportions in the opponent's 30-m zone; the purple markers show the ball control time points.

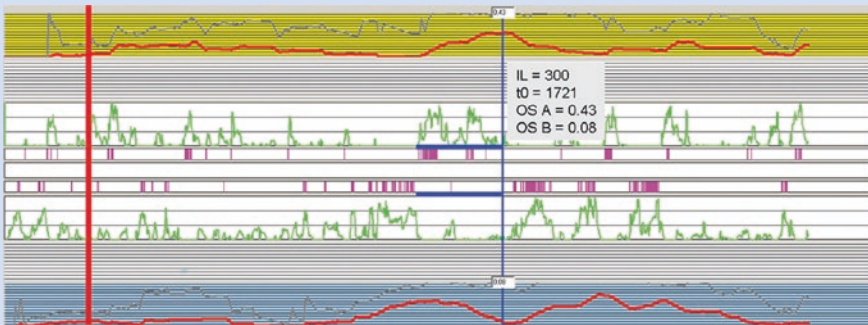


Fig. 8.5 Progressions of the attack efficiencies for the interval length $IL = 300$

8.3 Application

► Example 1

Physiological models for the optimization of stress-performance interactions are used to simulate

- Short-term effects of competition load on performance and fatigue;
- Long-term effects of training load on performance and recovery requirements (Chap. 13).

The central idea of modeling is to reduce the complex physiological interactions to the essential aspects of stress and performance. In this context, the delays with which

load and recovery take effect are the focus of attention: the shorter the recovery delay compared to the load delay, the more developed performance and capability are. Based on these analysis data, training and competition can be improved in their effect (Tampier et al., 2012). If the data expected from the analysis does not match the measured data, this may indicate an irregular training situation such as an unrecognized illness or illicit aids (e.g., doping). ◀

▶ Example 2

Tactical-strategic models to represent and analyze player behavior in team games have been used in soccer:

Formations: The player distribution of a team or its tactical arrangements can be analyzed by artificial neural networks and thus reduced to a few prototypical formations (Grunz et al., 2012; Perl et al., 2013). With the help of a simulative dynamics analysis of formation changes in specific game situations, tactical behavior patterns can be identified and then, for example, optimized, avoided, or disrupted (the opponent) (Perl & Memmert, 2017). In this context, the modeling of creativity or creative solutions in sports play is also successful (Memmert & Perl, 2009a, 2009b).

Voronoi cells: As shown above, Voronoi cells help to analyze the spatial control of players, teams, or tactical groups. Together with ball control, which can be analyzed from the position and movement data of players and ball, one can thus develop models that calculate the efficiency of attacking behavior from the coincidence of space and ball control relative to the players' action effort (Perl & Memmert, 2015). ◀

? Questions for the Students

1. How could a physiological stress-performance model be used to identify a doping violation?
2. How accurate must a Voronoi model of a soccer match be?
 - (a) Representation precision in video standard?
 - (b) Approximately, and only for critical phases?
 - (c) In coordination with the analysis requirements?

References

-
- Grunz, A., Memmert, D., & Perl, J. (2012). Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human Movement Science, 31*, 334–343.
- Low, B., Coutinho, D., Gonçalves, B., Rein, R., Memmert, D., & Sampaio, J. (2019). A systematic review of collective tactical behaviours in football using positional data. *Sports Medicine, 50*, 343–385.
- Memmert, D. (Ed.). (2021). *Match analysis*. Routledge.
- Memmert, D., Lemmink, K., & Sampaio, J. (2017). Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine, 47*, 1–10.
- Memmert, D., & Perl, J. (2009a). Analysis and simulation of creativity learning by means of artificial neural networks. *Human Movement Science, 28*, 263–282.
- Memmert, D., & Perl, J. (2009b). Game creativity analysis by means of neural networks. *Journal of Sport Science, 27*, 139–149.

- Memmert, D., & Raabe, D. (2018). *Data analytics in football. Positional data collection, modelling and analysis*. Routledge.
- Perl, J. (2015). Modelling and simulation. In A. Baca (Ed.), *Computer science in sport* (pp. 110–153). Routledge.
- Perl, J., Grunz, A., & Memmert, D. (2013). Tactics in soccer: An advanced approach. *International Journal of Computer Science in Sport*, 12, 33–44.
- Perl, J. & Memmert, D. (2015). Analysis of process dynamics in soccer by means of artificial neural networks and Voronoi-cells. In A. Baca & M. Stöckl (eds.), *Schriften der Deutschen Vereinigung für Sportwissenschaft, Band 244*, (S. 130–135). Hamburg: Czwalina.
- Perl, J., & Memmert, D. (2017). A pilot study on offensive success in soccer based on space and ball control—key performance indicators and key to understand game dynamics. *International Journal of Computer Science in Sport*, 16(1), 65–75.
- Perl, J., & Memmert, D. (2019). Soccer: Process and interaction. In A. Baca & J. Perl (Eds.), *Modeling and simulation in sport and exercise* (pp. 73–94). Routledge.
- Perl, J. & Uthmann, Th. (1997). Modellbildung. In J. Perl, M. Lames & W.-D. Miethling (Hrsg.), *Informatik im Sport. Ein Handbuch*. (pp. 65–80). Schorndorf 1997.
- Rein, R., Raabe, D., & Memmert, D. (2017). “Which pass is better?” Novel approaches to assess passing effectiveness in elite soccer. *Human movement science*, 55, 172–181.
- Tampier, M., Ender, S., Novatchkov, H., Baca, A., & Perl, J. (2012). Development of an intelligent real-time feedback system. *International Journal of Computer Science in Sport*, 11(3).



Predictive Models

Fabian Wunderlich

Contents

- 9.1 Example Sport – 66**
- 9.2 Background – 67**
 - 9.2.1 Looking into the Future – 67
 - 9.2.2 Predictive Models in Sports – 67
 - 9.2.3 Creation of Predictive Models – 68
 - 9.2.4 Exemplary Methods – 69
- 9.3 Applications – 70**
- References – 71**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Predictive models are relevant in almost all areas of science and society, including forecasts on weather and climate, politics and demography, as well as stock prices or economic growth.
- Driven by public and media interest, high data availability, and financial incentives, the domain of sports is a predestined environment for predictive models.
- Predictive models pursue a clearly defined goal, are based on suitable and meaningful data, use appropriate methodological approaches to statistical modelling, as well as suitable criteria to assess predictive quality.
- Methodologically, statistical models or computer science models (such as machine learning models) are suitable approaches. Among other things, the selection of methods depends on the exact problem definition and the data complexity.

9.1 Example Sport

Predictive thinking is a common approach in sports: Who will win the Super Bowl? Which youth player will have a successful career? What will be the results of the next Premier League matchday? Does the national team have a chance to win the Basketball World Cup? Which tactical formation is most promising in the upcoming match? How can injuries be avoided by adjusting training load? And who will win Wimbledon? Such questions are not only popular topics of conversation for the media and sports fans, but also form the basis of the bookmakers' business model and are important aspects in ensuring the sporting and financial success of sports teams. Thus, predictive models play a significant role in sports, in particular with regard to the interdisciplinary combination of computer science, mathematics, and sports science.

9.2 Background

9.2.1 Looking into the Future

Looking into the future seems to be part of normal human behaviour, especially in our modern society. We want to ensure survival by anticipating natural disasters or developments such as climate change. We want to live safely by anticipating threats from everyday crime to terrorist attacks. We want to secure our financial success by correctly anticipating stock market developments, economic trends, or consumer behaviour. And finally, we simply want to predict whether it will rain tomorrow so that we will have an umbrella with us when needed.

Forecasting is therefore a topic that is receiving attention in almost all areas of science and society. The domains where predictive models play an important role include economics (Timmermann, 2000), weather (Taylor & Buizza, 2004), climate (Green et al., 2009), political elections (Wolfers & Leigh, 2002), political conflicts (Brandt et al., 2014), crime (Gorr et al., 2003), demography (Booth, 2006), or energy demand (Hong et al., 2016).

9.2.2 Predictive Models in Sports

Sports is another popular application field for predictive models (Horvat & Job, 2020; McHale & Swartz, 2019; Vaughan Williams & Stekler, 2010; Wunderlich & Memmert, 2020), whose relevance is supported by the specific characteristics of sport.

Due to high media and public interest, there is a large amount of available data, allowing predictive models to draw on datasets with large sample sizes and/or a high level of detail (Angelini & de Angelis, 2019; Klemp et al., 2021; Koopman & Lit, 2019; Lessmann et al., 2010; Štrumbelj & Vračar, 2012).

In addition, there are several incentives for good predictive models. The sports betting market offers strong financial incentives for profitable forecasts of game outcomes, both on the side of bookmakers and professional bettors (Boshnakov et al., 2017; Constantinou et al., 2012; Hubáček et al., 2019). Moreover, the sports business itself offers high sporting and financial incentives, e.g., to adequately model spectator interest (Mueller, 2020; Van Reeth, 2019), optimal tactical movement behaviour on the field (Dick & Brefeld, 2019; Le et al., 2017), or risk of injury (Rossi et al., 2018).

Last but not least, there is a scientific interest as predictive models in sport can help to investigate general scientific theories and concepts such as market efficiency (Angelini & de Angelis, 2019; Bernardo et al., 2019; Direr, 2011; Goddard & Asimakopoulos, 2004) or crowd wisdom, i.e. collaborative human judgement (Peeters, 2018; Spann & Skiera, 2009).

9.2.3 Creation of Predictive Models

In this section, the necessary steps to create a predictive model are explained and methods from mathematics and computer science are highlighted based on two exemplary models.

Step 1: Goal

First of all, each predictive model is supposed to address a specific goal. This refers to solving one of the numerous application examples already mentioned in this chapter. Moreover, several further questions concerning the characteristics of the model have to be answered, e.g.

- Is a binary (yes/no) prediction or a percentage forecast needed?
- Should the model be simple and intuitive to understand?
- Is computation time influencing the value of the model?
- Does the model aim at high accuracy or high profitability?

Step 2: Data

Data is one of the most important and often limiting aspects of model selection. In particular, it is important to assess which data are available, whether they can be used freely and if so, in what quantity (sample size) the data are available. It is also important to consider that for meaningful predictive models, the data set should be divided into a sufficiently large training and test data set (in-sample and out-of-sample data). Furthermore, the application of a model only makes sense if data quality and information content are sufficient. Even the most sophisticated model will not be able to provide satisfactory results if the underlying data are incomplete, erroneous, or do not contain the required information.

Step 3: Methodological Approach

This step deals with the choice of the model itself, i.e. with the question of how the given data can be transferred into a forecast. We assume that a statistical or computer science approach is chosen for this purpose. Obviously, the approach taken should consider the goal of the model, data availability and complexity, as well as existing knowledge about state-of-the-art models or mechanisms underlying the processes. For example, machine learning models are particularly suitable in cases with a high level of data complexity, a lack of knowledge about the processes being modelled, and a low need for an intuitive understanding of the results.

Step 4: Evaluation of Predictive Quality

The final step is to define which criteria are suitable to assess the predictive quality of a model. This applies both to the calibration of a model and to the final evaluation of the predictive quality. Again, selection of the criteria depends essentially on the goal of the model. Common measures include the proportion of correct predictions, statistical measures of the accuracy of percentage forecasts, or profitability measures such as betting returns.

9.2.4 Exemplary Methods

The following two predictive models are intended to illustrate exemplary methods. Both are related to the outcome of sports events, and take methodologically different approaches borrowed from statistics as well as computer science.

Model 1: Statistical Model to Forecast Soccer Results (Hvattum & Arntzen, 2010)

The model introduced by Hvattum and Arntzen (2010) focuses percentage forecasts of results (home win, draw, away win) of soccer matches, primarily emphasizing the accuracy of the model. The data set consists of results of over 30,000 soccer matches, further match statistics are not used. Each team is assigned a parameter - the so-called ELO rating—which quantifies its playing strength. Prior to each match, an expected result is calculated based on these parameters. The strength parameters are then adjusted after each match based on the observed result. This yields an adaptive process in which the strengths of the teams receive continuous updates. To obtain a percentage forecast from the strengths of both teams, the authors use an ordinal logistic regression model. It receives the difference in strengths as input and determines the probabilities of home win, draw and away win.

Model 2: Computer Science Model for Forecasting Horse Racing (Lessmann et al., 2010)

The goal of the model by Lessmann et al. (2010) is to obtain percentage forecasts for horse races, primarily emphasizing the profitability of the model. The data set contains 1000 horse races and includes a wide variety of different parameters. In an attempt to find systematic factors influencing race results, the authors make use of a total of 41 different variables relating to the betting market, the circumstances of the race as well as the characteristics and prior results of horses, jockeys, and coaches. To convert these influencing factors into a percentage forecasts, the so-called Random Forest method is used. It belongs to the family of machine learning models, is based on the randomized generation of decision trees, and is described in more detail in a different chapter of this book.

Definition

In this chapter, predictive models particularly comprise all statistical or computer science models, that aim at estimating the probability of the occurrence of future events.

Study Box

Kovalchik (2016) analysed the quality of 11 different predictive models to forecast the outcomes of tennis matches. Based on over 2000 ATP singles matches from the 2014 season, she investigated regression models, point-based models, models based on pairwise comparisons and betting odds. Using four measures of predictive qual-

ity (prediction accuracy, calibration, log-loss and discrimination), she found that good models can predict the winner of a match in more than two-thirds of all cases. Regression models and ELO rating-based models performed best, while none of these mathematical models outperformed the predictive quality of betting odds. Moreover, the author reported that all models were more successful in forecasting matches of top players than matches of lower-ranked players.

9.3 Applications

► Example 1

This application area focuses probabilistic forecasts for the outcomes of sports events. Using soccer as an example, this includes the final result in terms of home win, draw, away win (Hvattum & Arntzen, 2010), the exact number of goals scored by both teams (Karlis & Ntzoufras, 2003), or the total number of goals in the match (Wheatcroft, 2020). This application example gains particular relevance from the possibility to bet on all these outcomes in the sports betting market. The data basis is usually prior results (Hvattum & Arntzen, 2010; Koopman & Lit, 2019) and/or additional team- or player-specific match statistics (Hubáček et al., 2019; Štrumbelj & Vračar, 2012). With regard to methodological approaches, classical statistical methods such as adaptive ratings and probability models are often used. ◀

► Example 2

This application area deals with predictive approaches for modelling the tactical behaviour of teams on the playing field. It is, therefore, more likely to be assigned to the areas of performance analysis or game analysis. Specifically, it attempts, for example, to forecast the collective movement behaviour of teams through so-called “ghosting” (Le et al., 2017; Seidl et al., 2018) or to analyse the dangerousness of game situations and actions on the field (Dick & Brefeld, 2019; Link et al., 2016; Lucey et al., 2014; Wei et al., 2013). Such approaches usually draw on extensive data sets of positional data and event data, which supports the use of machine learning models. ◀

► Example 3

A relatively recent application example is predictive approaches for injury prevention (Rossi et al., 2018) based on motion data, which establish a link to the fields of load control and sports medicine. Data basis can be, e.g., GPS motion data, further physical load data, and, if available, additional personal and medical data (Ehrmann et al., 2016; Rossi et al., 2018). Again, due to the multitude and complexity of data sources, machine learning models may be the most suitable approach (Claudino et al., 2019). ◀

🔍 Questions for the Students

1. What are important applications for predictive models in sports?
2. Which steps are necessary to develop a useful predictive model?

References

- Angelini, G., & de Angelis, L. (2019). Efficiency of online football betting markets. *International Journal of Forecasting*, 35(2), 712–721. <https://doi.org/10.1016/j.ijforecast.2018.07.008>
- Bernardo, G., Ruberti, M., & Verona, R. (2019). Semi-strong inefficiency in the fixed odds betting market: Underestimating the positive impact of head coach replacement in the main European soccer leagues. *The Quarterly Review of Economics and Finance*, 71, 239–246. <https://doi.org/10.1016/j.qref.2018.08.007>
- Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 22(3), 547–581. <https://doi.org/10.1016/j.ijforecast.2006.04.001>
- Boshnakov, G., Kharrat, T., & McHale, I. G. (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2), 458–466. <https://doi.org/10.1016/j.ijforecast.2016.11.006>
- Brandt, P. T., Freeman, J. R., & Schrodt, P. A. (2014). Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting*, 30(4), 944–962. <https://doi.org/10.1016/j.ijforecast.2014.03.014>
- Claudino, J. G., Capanema, D. D. O., de Souza, T. V., Serrão, J. C., Machado Pereira, A. C., & Nassis, G. P. (2019). Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: A systematic review. *Sports Medicine—Open*, 5(1), 28. <https://doi.org/10.1186/s40798-019-0202-3>
- Constantinou, A. C., Fenton, N. E., & Neil, M. (2012). Pi-football: A Bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, 36, 322–339. <https://doi.org/10.1016/j.knosys.2012.07.008>
- Dick, U., & Brefeld, U. (2019). Learning to rate player positioning in soccer. *Big Data*, 7(1), 71–82. <https://doi.org/10.1089/big.2018.0054>
- Direr, A. (2011). Are betting markets efficient? Evidence from European Football Championships. *Applied Economics*, 45(3), 343–356. <https://doi.org/10.1080/00036846.2011.602010>
- Ehrmann, F. E., Duncan, C. S., Sindhusake, D., Franzsen, W. N., & Greene, D. A. (2016). Gps and injury prevention in professional soccer. *Journal of Strength and Conditioning Research*, 30(2), 360–367. <https://doi.org/10.1519/JSC.0000000000001093>
- Goddard, J., & Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1), 51–66. <https://doi.org/10.1002/for.877>
- Gorr, W., Olligschlaeger, A., & Thompson, Y. (2003). Short-term forecasting of crime. *International Journal of Forecasting*, 19(4), 579–594. [https://doi.org/10.1016/S0169-2070\(03\)00092-X](https://doi.org/10.1016/S0169-2070(03)00092-X)
- Green, K. C., Armstrong, J. S., & Soon, W. (2009). Validity of climate change forecasting for public policy decision making. *International Journal of Forecasting*, 25(4), 826–832. <https://doi.org/10.1016/j.ijforecast.2009.05.011>
- Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, 32(3), 896–913. <https://doi.org/10.1016/j.ijforecast.2016.02.001>
- Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *WIREs Data Mining and Knowledge Discovery*, 10(5). <https://doi.org/10.1002/widm.1380>
- Hubáček, O., Šourek, G., & Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, 35(2), 783–796. <https://doi.org/10.1016/j.ijforecast.2019.01.001>
- Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3), 460–470. <https://doi.org/10.1016/j.ijforecast.2009.10.002>
- Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (the Statistician)*, 52(3), 381–393.
- Klemp, M., Wunderlich, F., & Memmert, D. (2021). In-play forecasting in football using event and positional data. *Scientific Reports*, 11(1), 24,139. <https://doi.org/10.1038/s41598-021-03157-3>

- Koopman, S. J., & Lit, R. (2019). Forecasting football match results in national league competitions using score-driven time series models. *International Journal of Forecasting*, 35(2), 797–809. <https://doi.org/10.1016/j.ijforecast.2018.10.011>
- Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, 12(3), 127–138. <https://doi.org/10.1515/jqas-2015-0059>
- Le, H., Carr, P., Yue, Y., & Lucey, P. (2017). Data-driven ghosting using deep imitation learning. In *Proceedings of the 11th annual MIT Sloan sports analytics conference 2017*. Boston, MA.
- Lessmann, S., Sung, M.-C., & Johnson, J. E. (2010). Alternative methods of predicting competitive events: An application in horserace betting markets. *International Journal of Forecasting*, 26(3), 518–536. <https://doi.org/10.1016/j.ijforecast.2009.12.013>
- Link, D., Lang, S., & Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLoS One*, 11(12), e0168768. <https://doi.org/10.1371/journal.pone.0168768>
- Lucey, P., Bialkowski, A., Carr, P., Yue, Y., & Matthews, I. (2014). How to get an open shot: Analyzing team movement in basketball using tracking data. In *Proceedings of the 8th annual MIT SLOAN sports analytics conference*. Symposium conducted at the meeting of Citeseer.
- McHale, I., & Swartz, T. (2019). Editorial: Forecasting in sports. *International Journal of Forecasting*, 35(2), 710–711. <https://doi.org/10.1016/j.ijforecast.2019.01.002>
- Mueller, S. Q. (2020). Pre- and within-season attendance forecasting in Major League Baseball: A random forest approach. *Applied Economics*, 52(41), 4512–4528. <https://doi.org/10.1080/00036846.2020.1736502>
- Peeters, T. (2018). Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*, 34(1), 17–29. <https://doi.org/10.1016/j.ijforecast.2017.08.002>
- Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS One*, 13(7), e0201264. <https://doi.org/10.1371/journal.pone.0201264>
- Seidl, T., Cherukumudi, A., Hartnett, A., Carr, P., & Lucey, P. (2018). Bhostgusters: Realtime interactive play sketching with synthesized nba defenses. In *12th Annual MIT Sloan Sports Analytics Conference*.
- Spann, M., & Skiera, B. (2009). Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28(1), 55–72. <https://doi.org/10.1002/for.1091>
- Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting*, 28(2), 532–542. <https://doi.org/10.1016/j.ijforecast.2011.01.004>
- Taylor, J. W., & Buizza, R. (2004). A comparison of temperature density forecasts from GARCH and atmospheric models. *Journal of Forecasting*, 23(5), 337–355. <https://doi.org/10.1002/for.917>
- Timmermann, A. (2000). Density forecasting in economics and finance. *Journal of Forecasting*, 19(4), 231–234.
- Van Reeth, D. (2019). Forecasting Tour de France TV audiences: A multi-country analysis. *International Journal of Forecasting*, 35(2), 810–821. <https://doi.org/10.1016/j.ijforecast.2018.06.003>
- Vaughan Williams, L., & Stekler, H. O. (2010). Sports forecasting. *International Journal of Forecasting*, 26(3), 445–447. <https://doi.org/10.1016/j.ijforecast.2009.12.005>
- Wei, X., Lucey, P., Morgan, S., & Sridharan, S. (2013). Sweet-spot: Using spatiotemporal data to discover and predict shots in tennis. In *7th Annual MIT Sloan sports analytics conference*, Boston, MA.
- Wheatcroft, E. (2020). A profitable model for predicting the over/under market in football. *International Journal of Forecasting*, 36(3), 916–932. <https://doi.org/10.1016/j.ijforecast.2019.11.001>
- Wolfers, J., & Leigh, A. (2002). Three tools for forecasting federal elections: Lessons from 2001. *Australian Journal of Political Science*, 37(2), 223–240. <https://doi.org/10.1080/10361140220148115>
- Wunderlich, F., & Memmert, D. (2020). Forecasting the outcomes of sports events: A review. *European Journal of Sport Science*, 21(7), 944–957. <https://doi.org/10.1080/17461391.2020.1793002>



Physiological Modeling

Manuel Bassek

Contents

10.1 Example Sport – 74

10.2 Background – 75

10.3 Applications – 76

References – 78

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Physiological models are used to describe the demands of physical activity and analyze the load induced on athletes in their sport.
- The collection of position data gives the opportunity to calculate the objective external load precisely
- Physiological models can be used to extrapolate from external load to internal load.
- Different physiological models are suited for different activities, like constant speed running in track and field or non-linear movement in handball.

10.1 Example Sport

This chapter introduces three different approaches to physiological modeling of player load from objective data sources, like position data. Coaches can profit from such data, especially when dealing with large numbers of players, like in team sports. Here, monitoring player load is a central aspect for coaches during training and competition. Accurate information about the stress and regeneration imposed on an athlete can help to constantly push performance levels to the maximum while minimizing the risk of injuries. It can be used to control adherence to prescribed training, analyze if specific drills had the desired physiological effect, plan individual regeneration periods or monitor recovery processes after injuries. The central advantages of physiological modeling are, on the one hand, the instantaneous evaluation of player load, as modern algorithms can evaluate drills, training sessions or matches in split seconds. On the other hand, it allows decision making during training to be based on objective data and empirical evidence.

10.2 Background

To optimize training and prevent injuries, accurate characterization of player load is an important task for coaches, match analysts and researchers (Akenhead & Nassis, 2016; Bourdon et al., 2017). Player load is generally defined as the relative biological stressors imposed on the athlete during training or competition (Bourdon et al., 2017). A stressor can be anything in training that induces some biological response from the body while relative means, that the response to an objective stress depends on the individual capacities of the athlete. Those stressors can be either measured internal (e.g., heart rate) or external (e.g., running distance) and obtained subjectively (e.g., rate of perceived exertion) or objectively (GPS data). Although objective and internal measures (e.g. rate of oxygen consumption) would be favorable to assess player load, they are usually not suitable to use outside of a laboratory setting.

Regardless, individuals can already measure objective internal and external load with widely available wearables, like heart rate monitors, GPS capable watches or smart phones (Lutz et al., 2019). The latter give accurate measures on distances, velocities, and accelerations (Scott et al., 2016). In professional team sports, like football, collection of position data is a standard procedure during practice and competition (s. Chap. 6) and makes it possible to monitor players activity continuously. However, the raw data is not comprehensible for analysts to derive adequate training recommendations. For example, the recording of one full handball match with a local position system will result in over a million data points (14 players \times 60 min \times 60 s \times 20 Hz). Therefore, data has to be condensed into interpretable and comparable measures of player load to be able to track inter- and intraindividual work loads effectively. Remember that player load refers to *relative* and *biological* stressors. Thus, physiological modeling of player load aims at extracting parameters from objective external measures that are closely related to objective internal measures. Ideally, those can be put into context of the athlete's individual capacities.

Definition

Physiological modeling of player load describes the extraction of physiological parameters from non-physiological data. It can be used to monitor athletes workloads during training and competition to optimize training results and prevent injuries (Akenhead & Nassis, 2016; Bourdon et al., 2017). In professional team sports, the collected position data can be processed for that purpose.

Study Box

Bassek et al. (2023) analyzed the player load of elite handball players during 77 matches of the German Handball Bundesliga. They reported benchmark values for player load in elite handball matches such as distances in six speed zones, Metabolic power, Metabolic work, Equivalent distance and Equivalent distance index. Additionally, they compared the influence of the physiological model on the measured player load statistically. For that purpose, they calculated the distance covered and Equivalent distance for every player. The difference between them were then compared between the positions of wings, backcourts and pivots with an ANOVA. The results show a significant

interaction effect for the difference between distance and Equivalent distance and the player positions. Wings had a larger difference between distance and Equivalent distance than backcourts and pivots. This means that the game of wings is more characterized by frequent accelerations and decelerations. The results are in line with other studies that report that wings were more frequently involved in counter attacks which require maximum accelerations. The choice of the right physiological model is therefore crucial for the analysis of player load. For sports that are characterized by accelerations and decelerations models that implement them should be used.

10

10.3 Applications**► Example 1**

Training impulse (TRIMP). Banister (1991) suggested the TRIMP to model internal load during endurance training. The TRIMP is calculated as the product of training duration and intensity. The intensity is derived from heart rate data and modeled in an exponential function to account for the non-linear relationship between intensity and load, as seen by blood lactate curves (for detailed formula, see Borresen & Lambert, 2009). It further includes the resting and maximum heart rate as representations of the athlete's individual condition. Even individual lactate curves can be included in the model (Manzi et al., 2014). The TRIMP allows for ecologic measurement of player load as it only requires heart rate measurements. It can be used to monitor the intensity of individual sessions of prolonged endurance training. In team sports, it can be applied during (pre-)season conditioning and regeneration. However, it is limited to constant speed sessions and can not be used during interval training or sport specific drills. ◀

► Example 2

Speed zones. Multidirectional team sports, like handball are characterized by the non-linear movement behavior of players, which means that players are constantly changing their speed (Karcher & Buchheit, 2014). One approach to measure player load in team sports is to divide the distance covered by players into speed zones. Over the time, many

models with different numbers of zones and cut-off speed have been used (see Miguel et al., 2021 for a detailed review). For example, Aslan and Aç (2012) used 8 zones: (1) walking: 0–6 km/h, (2) jogging: 6.1–8 km/h, (3) low-intensity running: 8.1–12 km/h, (4) moderate-intensity running: 12.1–15 km/h, (5) high-intensity running: 15.1–18 km/h, (6) low-intensity string: 18.1–21 km/h, (7) moderate-intensity sprint: 21.1–24 km/h, (8) high intensity sprint: > 24 km/h; whereas Clemente et al. (2019) identified 4 zones: (1) walking: 0–6.9 km/h, (2) jogging: 7–13.9 km/h, (3) running: 14–20 km/h, (4) sprinting: >20 km/h. Other approaches individualize the speed zones based on athletes capacities measured by lactate thresholds (Aslan & Aç, 2012) or in percentages of the maximum speed (Bacon & Mauger, 2017).

Distances covered in different speed zones is a simple approach to modeling player load when the activity is not linear in nature. It can be especially useful to compare the distance-per-speed zone profiles of athletes with desirable benchmark values. Such values can be the own performance prior an injury to compare during rehabilitation or the average professional player to identify talents. However, the different definitions of speed zones make it difficult to compare models used by different researchers and practitioners (Bradley & Ade, 2018). ◀

► Example 3

Metabolic power. Both approaches described previously do not incorporate accelerations and decelerations in their modeling. A possible way to include them is the concept of metabolic power. Metabolic power is defined as the energy expenditure per unit of time necessary to move at a certain speed, and is calculated as the product of energy cost of transport, per unit body mass and distance ($\text{J}\cdot\text{kg}^{-1}\cdot\text{m}^{-1}$) and velocity ($\text{m}\cdot\text{s}^{-1}$) (di Prampero & Osgnach, 2018). The measure was first introduced by di Prampero et al. (2005), who used the biomechanical equivalence of accelerated (or decelerated) running on flat terrain and constant running uphill (or downhill) to estimate the energy requirement for a specific displacement. Since then it has been used in several studies to characterize player load (Miguel et al., 2021).

The metabolic power model can provide several parameters of player load. (1) Metabolic power: The instantaneous power needed to perform the current locomotion, (2) Metabolic work: The energy needed to perform the locomotion in a time window, (3) Equivalent distance: The distance someone could have covered with the same energy if they didn't perform any accelerations or decelerations, (4) Equivalent distance index: The ration of equivalent distance and actual distance covered. The equivalent distance index is can be used as an indicator for how much an activity was characterized by accelerations and decelerations. These measures give a more detailed view into the internal load based on the combined analysis of distance, velocity and acceleration. Additionally, they provide comprehensive and comparable values of player load (Polglaze & Hoppe, 2019). ◀

? Questions for the Students

1. In which categories can the assessment of player load be divided?
2. Name three key measures that can be derived from position data and build the foundation for physiological modeling.

References

- Akenhead, R., & Nassis, G. P. (2016). Training load and player monitoring in high-level football: Current practice and perceptions. *International Journal of Sports Physiology and Performance*, *11*(5), 587–593. <https://doi.org/10.1123/ijsp.2015-0331>
- Aslan, A., & Aç, C. (2012). Metabolic demands of match performance in young soccer players. *Journal of Sports Science & Medicine*, *11*, 170–179.
- Bacon, C. S., & Mauger, A. R. (2017). Prediction of overuse injuries in professional U18-U21 footballers using metrics of training distance and intensity. *Journal of Strength and Conditioning Research*, *31*(11), 3067–3076. <https://doi.org/10.1519/JSC.0000000000001744>
- Banister, E. W. (1991). Modeling elite athletic performance. In D. MacDougall, H. A. Wenger, & H. J. Green (Eds.), *Physiological testing of the high-performance athlete* (2nd ed.). Human Kinetics Books.
- Bassek, M., Raabe, D., Memmert, D., & Rein, R. (2023). Analysis of motion characteristics and metabolic power in elite male handball players. *Journal of Sports Science and Medicine*, 310–316. <https://doi.org/10.52082/jssm.2023.310>
- Borresen, J., & Lambert, M. I. (2009). The quantification of training load, the training response and the effect on performance. *Sports Medicine*, *39*(9), 779–795.
- Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., Gabbett, T. J., Coutts, A. J., Burgess, D. J., Gregson, W., & Cable, N. T. (2017). Monitoring athlete training loads: Consensus statement. *International Journal of Sports Physiology and Performance*, *12*(s2), S2-161–S2-170. <https://doi.org/10.1123/IJSP.2017-0208>
- Bradley, P. S., & Ade, J. D. (2018). Are current physical match performance metrics in elite soccer fit for purpose or is the adoption of an integrated approach needed? *International Journal of Sports Physiology and Performance*, *13*(5), 656–664. <https://doi.org/10.1123/ijsp.2017-0433>
- Clemente, F. M., Owen, A., Serra-Olivares, J., & Nikolaidis, P. T. (2019). Characterization of the weekly external load profile of professional soccer teams from Portugal and The Netherlands. *Journal of Human Kinetics*, *66*, 155–164. <https://doi.org/10.2478/hukin-2018-0054>
- di Prampero, P. E., Fusi, S., Sepulcri, L., Morin, J. B., Belli, A., & Antonutto, G. (2005). Sprint running: A new energetic approach. *Journal of Experimental Biology*, *208*(14), 2809–2816. <https://doi.org/10.1242/jeb.01700>
- di Prampero, P. E., & Osgnach, C. (2018). Metabolic power in team sports—part 1: An update. *International Journal of Sports Medicine*, *39*(08), 581–587. <https://doi.org/10.1055/a-0592-7660>
- Karcher, C., & Buchheit, M. (2014). On-court demands of elite handball, with special reference to playing positions. *Sports Medicine*, *44*(6), 797–814. <https://doi.org/10.1007/s40279-014-0164-z>
- Lutz, J., Memmert, D., Raabe, D., Dornberger, R., & Donath, L. (2019). Wearables for integrative performance and tactic analyses: Opportunities, challenges, and future directions. *International Journal of Environmental Research and Public Health*, *17*(1), 1–26. <https://doi.org/10.3390/ijerph17010059>
- Manzi, V., Impellizzeri, F., & Castagna, C. (2014). Aerobic fitness ecological validity in elite soccer players: A metabolic power approach. *Journal of Strength and Conditioning Research*, *28*(4), 6–919.
- Miguel, M., Oliveira, R., Loureiro, N., García-Rubio, J., & Ibáñez, S. J. (2021). Load measures in training/match monitoring in soccer: A systematic review. *International Journal of Environmental Research and Public Health*, *18*(5), 2721. <https://doi.org/10.3390/ijerph18052721>
- Polglaze, T., & Hoppe, M. W. (2019). Metabolic power: A step in the right direction for team sports. *International Journal of Sports Physiology and Performance*, *14*(3), 407–411. <https://doi.org/10.1123/ijsp.2018-0661>
- Scott, M. T. U., Scott, T. J., & Kelly, V. G. (2016). The validity and reliability of global positioning systems in team sport: A brief review. *Journal of Strength and Conditioning Research*, *30*(5), 1470–1490. <https://doi.org/10.1519/JSC.0000000000001221>

Simulation

Contents

- Chapter 11 Simulation – 81**
Jürgen Perl and Daniel Memmert
- Chapter 12 Metabolic Simulation – 89**
Dietmar Saupe
- Chapter 13 Simulation of Physiological
Adaptation Processes – 99**
Marc Pfeiffer and Stefan Endler



Simulation

Jürgen Perl and Daniel Memmert

Contents

11.1 Example Sport – 82

11.2 Background – 83

11.3 Applications – 86

References – 88

Jürgen Perl was deceased at the time of publication.

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com.

Key Messages

System—Model—Simulation:

- System behavior becomes computable by models;
- Model calculations simulate system behavior and make it transparent;
- Simulations help to understand, predict and influence system behavior.

11.1 Example Sport

11

Performance analysis in sports has two main goals: On the one hand, the recognition and optimal use of individual performance limits in competition, especially during continuous stress. On the other hand, the achievement and, if necessary, increase of physiological performance capacities in training. Due to the complex interactions of the most different physiological components of the body, it is difficult to predict effects like under- or overload. An interaction model simplified to a few parameters makes it possible, by varying these parameters (in PerPot (cf. Chap. 13): Effect Delays) to simulate the physiological responses and thus optimize them. The method of regular review and concluding optimization, which is adequate for training, can then be applied accordingly in competition through on-line data collection and optimization. Beyond this optimization, however, discrepancies between expectation and reality can also be detected by such simulation-based analyses: On the one hand, the actual performance course may be significantly below the simulated course and thus signal an additional physiological stress situation (e.g., illness). On the other hand, the performance course could be significantly above the expectation, which would then suggest additional performance potentials (such as doping).

11.2 Background

Simulation-based approaches to behavioral optimization can be found in sports in areas as diverse as tactics optimization in team games, technique optimization for movement patterns, or load-performance optimization in training and competition. For the last example, the connection between reality, model, and simulation can be illustrated in a particularly vivid way. Physiological performance analyses can be performed with *PerPot* (cf. Perl, 2002, 2004), which has been used in various fields of performance analysis and optimization (see Fig. 11.1). According to the explanations in Chap. XX, in the *PerPot* model the human physiological system is reduced to a minimum of (abstract) basic components, whose interactions nevertheless reflect with astonishing precision the stress-performance dynamics (not only) in sports and is thus suitable for prognostic simulations (cf. Perl, 2003):

A loading rate (e.g. running speed) fills, in the same way, the two internal potentials “load” and “recovery”. From the potential “recovery” the power potential (here e.g. heart rate) is filled with a delay DR (delay in response). From the potential “load”, power is drawn off with a delay DS (delay in strain) via a negative flow. This results in a transient process on the power potential, which may lead to stable conditions or—e.g. in case of overloads—to complete exhaustion of the power potential. In the load potential, the “reserve” is an indicator of how high the load level is. If the load potential overflows due to a very high load rate, i.e. the reserve becomes negative, then with a very short delay DSO (delay in strain overflow), the power potential is additionally reduced. The case of negative reserve can be used as an overload indicator.

In the following figures, the top graphs each show an equal load profile (running speed), the middle graphs show power profiles (heart rate), and the bottom graphs show the reserve profiles as indicators of the system state.

Fig. 11.1 The *PerPot* performance potential model. (After Perl, 2003)

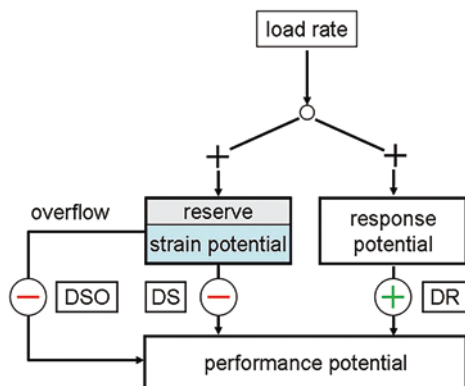


Figure 11.2 shows how the calculated heart rates (green, directly proportional) and the reserve (gray, inversely proportional) track the velocity profile with slight delays. These delays and thus the overall course are essentially caused by the recovery delay DR: The smaller DR, i.e. the shorter the recovery delay, the lower and smoother the heart rate course, and the higher the reserve values.

Two essential aspects and applications for power simulation are:

1. the detection of inadequate load patterns (overload, underload) and the simulative optimization of load-power dynamics.

To analyze athletic performance, e.g., in terms of potential for improvement, one can vary the delay parameters, in the following specifically the recovery delay DR, and thus identify the potential for increase in performance in the simulation (cf. Perl, 2003). Figure 11.3 shows the heart rate curve (green) and the reserve curve (gray) of the athlete from Figure 11.2 with a recovery delay of $DR = 7.2$. Supplementary simulated as a training target are the desired profiles (heart rate: gray, significantly lower and smoother; reserve: black, significantly higher) of a better-trained athlete, which can however only be achieved with a faster recovery, $DR = 5.8$.

The simulation thus allows to test and compare load profiles without overloading the athlete with too high training loads in real tests. In the example above, the result of the simulation would be the question of whether the DR value for the athlete under consideration could be reduced to below 6 at all by training. To answer this question, it would be useful to accompany the training sessions with appropriate simulations during the training process to detect and avoid possible overload situations in time.

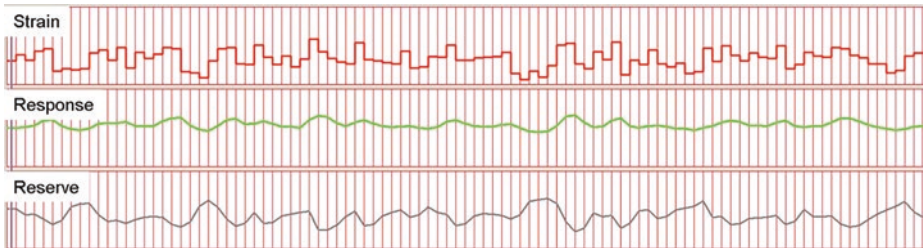


Figure 11.2 Standard simulative calculation

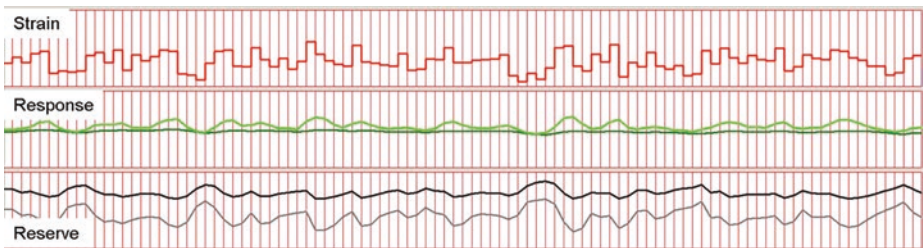
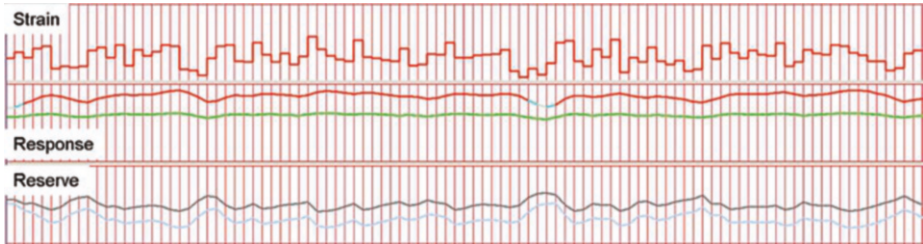


Figure 11.3 Comparison of actual and target profiles



■ Fig. 11.4 Comparison of expected and actual profiles

2. the detection of implausible load-performance dynamics, caused e.g. by conspicuous deceleration values, which in turn indicate physiological manipulations (e.g. doping).

Example: In ■ Fig. 11.4, the simulated heart rate progression of an insufficiently trained athlete ($DR = 12.5$) is shown as a red profile in the middle graph: The pulse level is very high, the reserve values (gray) are rather low, and the reactions to load changes are strong. Overall, the expectations for a good competition result are very low.

In contrast, the profiles for heart rate (green) and reserve (black) recorded from the competition suggest an excellent performance level. A model-based analysis results—in stark contradiction to the preliminary analyses—in a very good value of $DR = 5.8$ for the recovery delay.

Definition

Simulation is the calculation of system behavior. The starting point for simulation is the model of a system, whose parameters and input data can be varied, and with the help of which calculations of the model behavior can be performed (Perl, 2015).

A simulation is used to

1. better understand the behavior of the system through parameter variation;
2. predict future behavior of the system;
3. detect anomalies in the behavior of the system.

Study Box

In individual sports, as described above, the focus is on simulating physiological performance. In contrast, in rebound or team games, the optimization of tactical behavior is the subject of simulation, as illustrated below using soccer as an example. The key to success lies in find-

ing the perfect mix of changing tactical patterns, which depend significantly on the behavior of the opposing team (Memmert & Raabe, 2018). Memmert et al. (2021) studied professional soccer matches according to the specific tactical team behavior “attack vs. defense” based

on a simulation approach. The formation patterns of all matches (40 positional datasets) are categorized by SOCCER© (Perl & Memmert, 2011) for defense and offense. Monte Carlo simulation can evaluate the mathematically optimal strategy. The interaction simulation between offense and defense shows optimal flexibility values for both tactical groups. The results showed that both offense and defense have optimal planning rates to be more successful. The more complex the success indicator, the more successful attacking player groups

become. The results also show that defensive player groups always succeed in attacking groups below a certain planning rate value. Simulation-based positional data analysis reveals successful strategic behavior patterns for attack and defense. Attacking player groups need very high flexibility (for creativity, see Memmert & Perl, 2009a, 2009b) to be successful (keep possession of the ball). Defensive player groups, on the other hand, only need to be below a defined flexibility level to guarantee more success.

11.3 Applications

► Example 1

Detection of possible overload phases in backstroke games like tennis are possible (unpublished analysis for EU project ► <https://matchpoint.bgtennis.bg/>). Stroke changes in backstroke games can be long and energy-consuming due to sprints to the receiving points. ■ Figure 11.5 shows on the left the motion profile of a tennis player and on the right the corresponding progressions of

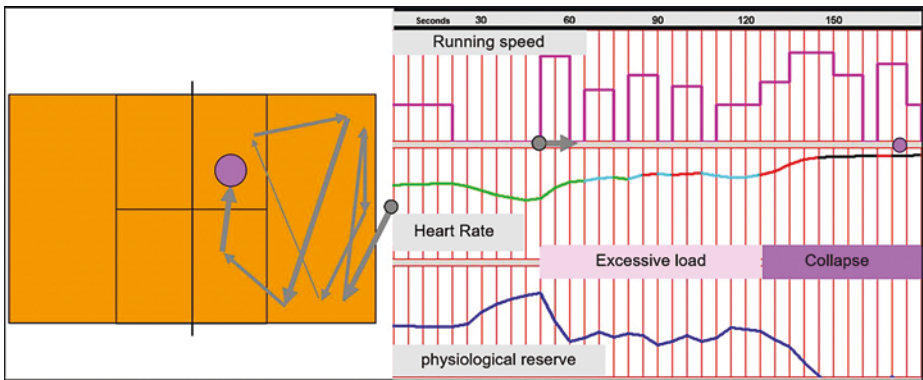
- Load: running speed with short pauses after own strokes;
- Power: heart rate, initially relaxed (green), then rising into the critical range (light blue, red) and finally ending in the collapse range (black);
- Reserve: in the first resting phase first strongly increasing, at the end in the overload phase decreasing into the negative collapse range.

These progressions can be diagnosed ex-post on the basis of recorded data or predicted by simulation ex-ante with the help of load-performance simulation for avoidance. ◀

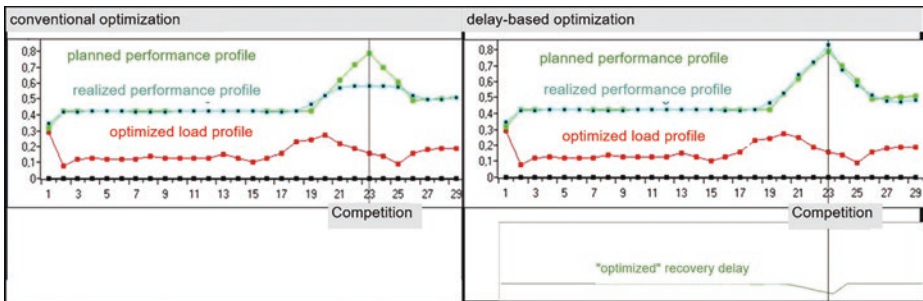
► Example 2

Recognition of implausible performance increases at the time of competition is possible (Perl, 2004). Peak loads in competition cannot be maintained continuously, but are prepared by a correspondingly building training program with an appropriate lead time. ◀

■ Figure 11.6 shows on the left a competition-oriented training profile (red), which causes an increase in performance (blue) at the competition date. However, the intended performance curve (green) could not be achieved. ■ Figure 11.6 shows the same initial situation on the right, but with an almost unchanged training profile, the planned peak performance is reached at the competition date. The reason for this is the recovery delay, which decreases significantly at the competi-



■ Fig. 11.5 Load development in a long fast stroke change in tennis



■ Fig. 11.6 Tactical measures for performance optimization

tion date and thus, as described above, significantly improves the ability to implement the load. Such a reduction of the recovery delay is, among other things, the goal of training, but usually works over weeks to months. A significant improvement within 3 days, on the other hand, is highly conspicuous and should be cause for a more detailed analysis of the measures taken.

► Example 3

Recognizing and optimizing strategies in team games where several players act “independently” of each other is difficult (Memmert, 2021). An innovative approach is to divide teams into a small number of tactical groups and analyze the interaction of these groups. The positions of players in tactical groups in soccer can then be mapped to formation patterns, reflecting strategic behavior and interaction (Perl & Memmert, 2019). Based on this information, Monte Carlo simulation allows generating tactical strategies that are optimal—at least from a mathematical point of view. In practice, behavior can be guided by these optimal strategies, but usually changes depending on the activities of the opposing team. Analyzing the game from the perspective of such simulated strategies can show how strictly or flexibly (cf. Memmert, 2015) a team varies strategic patterns. To optimize such team behavior of tactical group interactions in professional soccer, Perl et al. (2021) conducted a simulation and validation study based on 40 positional datasets from professional soccer using the SOCCER© software (Grunz et al., 2012; Perl et al., 2013; Perl & Memmert, 2011). After the validation study confirmed the applicability

of the defined tactical model, the simulation study showed that offensive player groups need less tactical flexibility to successfully gain possession of the ball, while defensive player groups need more tactical flexibility to do so. Offensive players should thus play with a more flexible tactical orientation to maintain possession, while defensive players should play with a more planned orientation to be successful. ◀

? Questions for the Students

1. Marathon running: Stressful running phases (e.g. bridges, cheering spectators) often have an effect only after a considerable delay. How could such stressful phases be recorded before or during the run and taken into account in (further) run planning?
2. Tactics optimization in soccer: The measurable positions of offensive and defensive players on the field change very quickly due to the situation (several thousand per player and half-time). How can you gather useful information about tactical concepts from this mass of data?

References

- Grunz, A., Memmert, D., & Perl, J. (2012). Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human Movement Science, 31*, 334–343.
- Memmert, D. (2015). *Teaching tactical creativity in sport: Research and practice*. Routledge.
- Memmert, D. (Ed.). (2021). *Match analysis*. Routledge.
- Memmert, D., Imkamp, J., & Perl, J. (2021). Flexible defends succeeds creative attacks!—A simulation approach based on position data in professional football. *Journal of Software Engineering and Applications, 14*(9). <https://doi.org/10.4236/jsea.2021.149029>
- Memmert, D., & Perl, J. (2009a). Analysis and simulation of creativity learning by means of artificial neural networks. *Human Movement Science, 28*, 263–282.
- Memmert, D., & Perl, J. (2009b). Game creativity analysis by means of neural networks. *Journal of Sport Science, 27*, 139–149.
- Memmert, D., & Raabe, D. (2018). *Data analytics in football. Positional data collection, modelling and analysis*. Routledge.
- Perl, J. (2002). Adaptation, antagonism, and system dynamics. In G. Ghent, D. Kluka, & D. Jones (Eds.), *Perspectives—The multidisciplinary series of physical education and sport science* (Vol. 4, pp. 105–125). Meyer & Meyer Sport.
- Perl, J. (2003). On the long-term behaviour of the performance-potential-metamodel PerPot: New results and approaches. *International Journal of Computer Science in Sport, 2*, 80–92.
- Perl, J. (2004). PerPot—A meta-model and software tool for analysis and optimisation of load-performance-interaction. *International Journal of Performance Analysis of Sport, 4*, 61–73.
- Perl, J. (2015). Modelling and simulation. In A. Baca (Ed.), *Computer science in sport* (pp. 110–153). Routledge.
- Perl, J., Grunz, A., & Memmert, D. (2013). Tactics in soccer: An advanced approach. *International Journal of Computer Science in Sport, 12*, 33–44.
- Perl, J., Imkamp, J., & Memmert, D. (2021). Key Strictness vs. flexibility: Simulation-based recognition of strategies and its success in soccer. *International Journal of Computer Science in Sport, 20*, 43–54.
- Perl, J., & Memmert, D. (2011). Net-based game analysis by means of the software tool SOCCER. *International Journal of Computer Science in Sport, 10*, 77–84.
- Perl, J., & Memmert, D. (2019). Soccer: Process and interaction. In A. Baca & J. Perl (Eds.), *Modelling and simulation in sport and exercise* (pp. 73–94). Routledge.



Metabolic Simulation

Dietmar Saupe

Contents

12.1 Example Sport – 90

12.2 Background – 91

12.3 Applications – 92

References – 97

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Mathematical modeling captures the functional relationship between the measurable metabolic output variables of lactate, oxygen uptake, carbon dioxide output, as well as heart rate, and the power demanded.
- Energy provision from the system of energy-rich phosphates and oxidative phosphorylation can also be done from a more complex meta-modeling as a system of coupled processes simulated on the computer with suitable algorithms.
- The (critical power) PC model describes the maximum achievable duration of a workload on the ergometer at a given power level.
- Models allow the estimation of interpretable parameters based on performance tests, the analysis of performance in training and competition, the monitoring of rehabilitation measures, and the planning of strategies for the optimal use of performance potential.

12.1 Example Sport

Two methodologies will be used to discuss how the dynamics and limits of metabolic energy provision can be described quantitatively through modeling, simulation, and analysis in sports informatics and used for practical applications. The examples are primarily related to road cycling and cycling ergometer tests, but can be adapted to other endurance sports such as running, swimming or rowing by appropriate modifications. In performance diagnostics, exercise science and sports medicine, methods are used to model and predict the effects of variable exercise demand on measurable indicators such as heart rate, oxygen uptake and lactate production. For athletes, this can yield valuable conclusions about fitness parameters and training success. In competition, it is important to use the individually

available energy supply in the best possible way. In every phase of a race, athletes must be able to estimate how much power they can produce without exhausting themselves prematurely, but also without arriving at the finish line with unused energy reserves. In road cycling, they can do this by using their measured power in watts and their heart rate as a guide, based on their years of training and competition experience. However, it is not enough to simply set a suitable power and maintain it consistently. Research using a theoretical approach has shown that on courses with variable gradient profiles or changing wind conditions, a variable power distribution is advantageous over a constant one. Mathematical modeling and simulation allows to develop appropriate adaptive pacing strategies.

12.2 Background

Performance diagnostics in endurance sports uses test procedures to quantitatively record the resilience and performance level of athletes. It provides a valuable basis for planning and controlling training. Of central importance is the part of the metabolism that generates the energy needed for the respective athletic load. Chemical reactions generate mechanical energy for the muscles.

The main energy source is adenosine triphosphate (ATP), which is only available in limited quantities in the muscle. Used ATP must be re-synthesized and the energy required for this is produced by oxidation of sugar (glycolysis), fats and proteins. This can be done aerobically or anaerobically, i.e. with or without the use of oxygen. In the case of anaerobic energy supply, a further distinction is made between lactic acid and alactic acid, i.e. with or without the production of lactic acid in the form of lactate. With increasing load, lactate is increasingly produced, which can no longer be broken down to the same extent. As a result, glycolysis is strongly inhibited, with the consequence that such high performance can no longer be maintained over the long term. The maximum power, indicated by speed in km/h on a treadmill or physical power in watts on an ergometer, at which lactate production and its breakdown are still in balance, is called the individual anaerobic threshold or maximum lactate steady state (MLSS).

Energy metabolism is thus a very complex network of many individual reactions. It is not possible to measure all the components directly, and only external indicators can provide indirect information about the current state of energy provision. The most important of these are lactate concentration in mmol/l from blood samples taken at the earlobe, oxygen uptake in ml/min measured by spiroergometry, heart rate in bpm, and finally, in cycling, the mechanical power generated in watts on the ergometer or in the field using power sensors on the crank, sprocket, pedals, or in the hub. A basic functionality of the models and simulations considered in this chapter is to determine the effects of power profiles (constant power, step test, intermittent training, or arbitrarily variable power in the field) on physiological measures.

Definition

Linear ordinary differential equations (ODE) with asymptotically constant solutions describe the physiological adaptation of oxygen uptake and heart rate to a constant power demand.

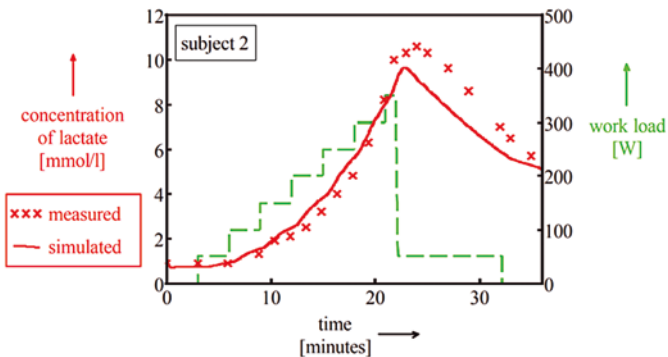
The (critical power) CP model describes the maximum amount of time that can be sustained at a given constant power on an ergometer to complete exhaustion.

Both models can be generalized for the case of a variable power demand.

12.3 Applications**► Example 1: Modeling and Simulation of Output Variables of Metabolic Processes**

Modeling and simulation are key technologies for understanding the behavior of such complex systems. An important requirement for these models for applications in endurance sports is to provide the measurable quantities mentioned as the result of a demand profile, the load. Based on comparison with data from studies on ergometers in the laboratory or measurements in the field, the parameters of the models can be estimated and the predictive power of the models can be assessed by simulation and comparison with the measured quantities. Basically, we can distinguish three approaches to modeling:

1. The metabolic processes are well known in biochemistry. For example, glycolysis can be divided into ten different chemical reactions. These convert glucose to pyruvate and release ATP in the process. From these and other reaction equations, a complex model consisting of differential equations and algebraic equations was set up and simulated in Schulte et al. (1999). Intervals for parameters could be taken in part from the literature on biochemistry and sports medicine, and the parameters themselves were determined iteratively by simulating the model using measurement data from laboratory experiments. As an example, Fig. 12.1 shows the comparison of lactate measurements with the lactate concentrations derived from the model.



■ Fig. 12.1 Step test on the ergometer until exhaustion and resulting blood lactate values from measurements and modeling, from Schulte et al. (1999)

This approach to modeling has the advantage that the variables and parameters have direct metabolic equivalents, although in many cases their sports science interpretation is not as obvious as with lactate. The disadvantage is the very high complexity, the required expert knowledge in biochemistry and in the mathematical methods for the setup and numerical solution of the resulting differential-algebraic equation system.

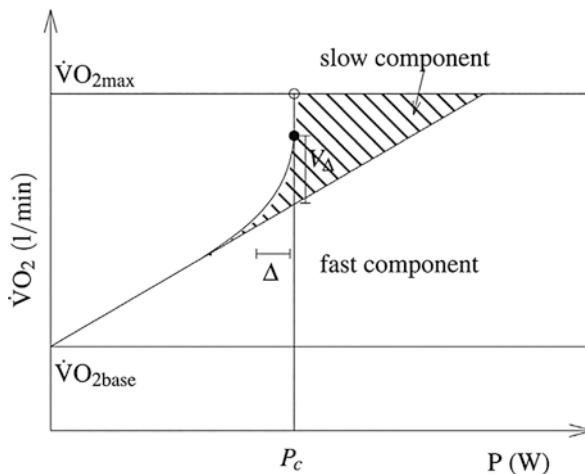
2. The much more common method of modeling directly sets up equations for the dynamics of the output variables oxygen uptake ($\dot{V}O_2$), lactate concentration ($[La]$), or heart rate (HR). Linear differential equations are preferred. These have exponential solutions that can well represent the adaptation of the metabolic system to a performance demand. The variables and parameters are easily interpretable, typically as amplitudes of adaptation responses and their associated time constants.

Fundamental to the dynamics of $\dot{V}O_2$ is first its value in equilibrium (steady-state), i.e. after completion of the adaptation response to a constant load set on the ergometer. This gives an individual monotonically increasing function of $\dot{V}O_2$ over a power interval from $P = 0$ (W) to a maximum and sustained critical power $P = P_c$ at which maximal oxygen uptake $\dot{V}O_{2max}$ is reached, see Jones and Poole (2013). The steady-state value of $\dot{V}O_2$ is essentially composed of three components:

(a) a baseline value that is slightly above the resting-state $\dot{V}O_2$, (b) a component A_1 that increases linearly with P , and (c) a smaller, so-called slow component A_2 that is added only above a certain power threshold, see [Fig. 12.2](#).

For both components A_1 and A_2 , the adaptation of $\dot{V}O_2$ to an incipient constant load P can be described very well by an exponential function with three parameters A , T and τ ,

$$A \left(1 - \exp \left(-\frac{t-T}{\tau} \right) \right)$$



[Fig. 12.2](#) Steady-state model for oxygen uptake $\dot{V}O_2$ as a function of load, a constant power P . Powers greater than the critical power P_c cannot be sustained. Figure from Artiga Gonzalez et al. (2019)

where A is the amplitude of the component in question in the steady-state and τ is a time constant quantifying the speed of adaptation to the steady-state. After τ time units, about 63% of the amplitude A is reached, after 3τ it is 95%. T is a time delay from which the component in question sets in.

These exponential dynamics can be described equivalently by the linear differential equation $\dot{x} = \tau^{-1}(A - x)$ with initial value $x(T) = 0$. This provides the approach to generalize the dynamic model for variable loads as they occur in training and competition in the field, i.e., for powers $P = P(t)$ that are not constant. For this purpose, the constant amplitude A in the ODE must be replaced by $A(P(t))$, i.e., by the amplitude given to the load P at time t according to the model in [Fig. 12.2](#). In Artiga Gonzalez et al. (2019) this was carried out and validated on experimental spiroergometric data series.

The same method can be used for modeling the dynamics of heart rate HR at variable load. In this case, the slow component can be omitted, so that the resting pulse HR_0 , the (constant) slope = dHR/dP (gain), the time constant τ , the time delay T and, if necessary, the critical power P_c or the maximum heart rate HR_{max} are sufficient as parameters.

In Mongin et al. (2020), this was used to fit the heart rate of 30 subjects during a treadmill exercise test. The power P was replaced by the speed of the treadmill. On median, 91% of the total variance of heart rate could be explained by the model. With an additional adjustment of the gain depending on the load intensity, this rate could even be improved to 99%. However, this requires a larger number of parameters to be estimated in the regression analysis and can easily lead to overfitting.

3. The above two modeling approaches aim to characterize the performance-related physiological response by explicit equations, based on the metabolic components and processes or the phenomenology of the measurable variables. In contrast, the relationship between dependent output (VO_2 , HR , $[La]$) and input (power or speed as load) is determined independently by black-box procedures. Machine learning methods such as support vector machines, neural networks, and deep learning have become widely used for this purpose (see Chaps. 20–24). These can of course also be applied to the analysis of metabolic data, see Zignoli et al. (2020). As expected, the generated systems for estimating output variables have a better fit compared to explicit modeling. However, the disadvantages are that significant amounts of data are required to train the neural networks and that the trained weight parameters do not give an obvious sports medicine interpretation.

For a more detailed introduction to the methods and background of modeling lactate and oxygen uptake in cycling, we recommend the recent paper by Zignoli et al. (2019). ◀

► Example 2: The (Critical Power) CP Model

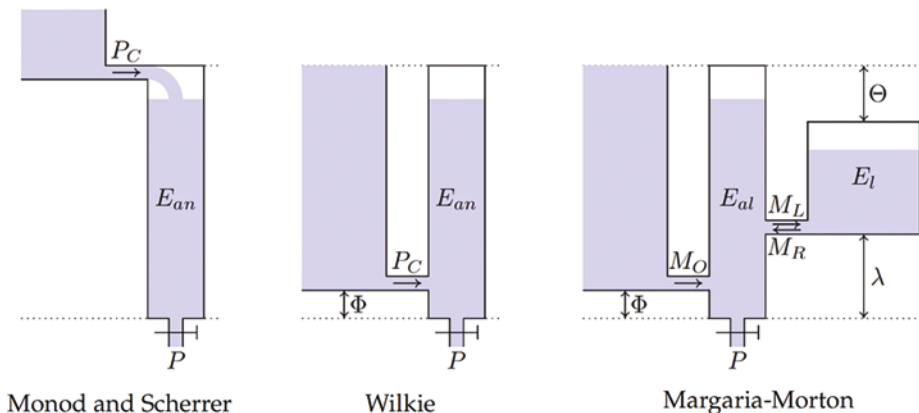
The methods discussed in the first example provide a basis for planning and controlling training, but in competition the main question for the cyclist is how much power he or she should aim for along the race course to reach the finish line as quickly as possible, especially if the course runs through variable terrain with different gradients. The answer requires knowledge of the maximum sustainable power P_c , the additional short-term

energy reserves available in each case for sprints, and the recovery rate at which the energy store can be replenished at more moderate power levels. A first mathematical model for this analysis has already been introduced more than 50 years ago by Monod and Scherrer (1965).

The basis of this concept is two different energy sources, aerobic and anaerobic. The aerobic energy supply is unlimited in magnitude, but can only be tapped at a fixed rate, the critical power P_c mentioned earlier. On the other hand, the anaerobic energy supply can be tapped at an unlimited rate, but its total magnitude E_{an} is quite limited. Consequently, an athlete exhausts himself after a certain time T , when the power P tapped is above the critical power P_c . In the nowadays so-called (critical power) CP model of Monod and Scherrer, the hyperbolic function $T = E_{an}/(P - P_c)$ is therefore applied for $P > P_c$. This approach can be well visualized as a hydraulic model (■ Fig. 12.3, left). Energy reservoirs are represented by tanks with water connected by pipes. The outflow below, controllable by a regulator, determines the requested power, given by the flow.


As an example, let's assume an amateur road cyclist with a critical power of 250 W and an anaerobic energy storage of 20,000 joules. How long can he ride at a power of 300 W? This gives $T = 20,000 \text{ Ws}/(300 \text{ W} - 250 \text{ W}) = 400 \text{ s}$, or 6 min and 40 s. A power of less than $P_c = 250 \text{ W}$, on the other hand, can be sustained indefinitely according to the model, provided enough food is consumed to maintain aerobic energy flow.

For constant load, the CP model dictates that the initial energy reserve E_{an} is consumed at a constant rate $P - P_c$ until it is exhausted after time T . To use the variable power demand model $P = P(t)$, we introduce the current energy reserve $e_{an}(t)$, with the default of the initial value $e_{an}(0) = E_{an}$. Consequently, this results in the differential equation $de_{an}/dt = P_c - P(t)$. This equation extends the model at the same time for the case of recovery at powers below P_c : each watt below P_c results in an inflow of 1 joule per second. ◀



■ **Fig. 12.3** Hydraulic representation of three physiological models. On the left, the classical model of Monod and Scherrer (1965); in the middle, the model of Wilkie (1981) with adaptive recovery rate; and on the right, that of Morton (1986) and Margaria with an additional vessel for anaerobic-lactacid energy provision. Figure from Wolf (2019)

The simple model for maximum sustainable power has worked well in practice. However, to determine the critical power P_c for an individual, one needs several points (P,T) on the graph of the hyperbolic function of T. For each of these points, an ergometer test must be performed to exhaustion, followed by a sufficiently long recovery period. This is hardly feasible in practice, and therefore simpler tests have been developed for this purpose, see the review in Lipková et al. (2022).

The 2-parameter CP model of Monod and Scherrer provided the foundation for a number of refinements to address minor shortcomings. It has been criticized that the model allows arbitrarily high power, at least for short periods, and that the direct update of the model estimates too optimistically for recoveries below critical power ($P < P_c$) as discussed above. In the 3-parameter model of Morton (1996), power is limited, and in Skiba et al. (2014) and Wolf (2019), the recovery rate was damped. Two other modifications of Wilkie, Margaria, and Morton are listed in  Fig. 12.3.

The physiological models combined with a mechanical model of required physical power during cycling on a track with a given elevation profile can be used to calculate best pacing strategies with numerical optimization, see Fayazi et al. (2013), Sundström and Bäckström (2017), and Wolf et al. (2019).

Study Box

Mathematical modeling of training and performance in sports informatics is a valuable tool for applications in coaching, fitness/personal training, rehabilitation, and exercise physiology. Two models, the Critical Power (CP) model and the Banister Impulse Response (IR) model, provide complementary methodologies for this purpose. The CP model describes the relationship between work performed and remaining energy reserves. The IR model describes the dynamics by which individual performance changes over time as a function of training. Both models elegantly abstract the underlying physiology. The comprehensive paper by Clarke and Skiba (2013) provides a detailed introduction of the related physiological principles, definitions, and history, derives the assumptions and equations, and instructs in the use of these resources using software (spreadsheets) for practical computer exercises.

12

Questions for the Students

1. Determine and discuss a model equation for heart rate at an onset constant power demand and the corresponding differential equation for variable power.
2. A test person performs two step tests with constant load on an ergometer. At 450 W he stops the test after only 1:20 min exhausted, after recovery he still manages 10:40 min at 275 W. What critical power P_c and total anaerobic energy E_{an} can be estimated from this? What values result if a third measurement at $P = 330$ W yields a total time of $T = 5$ min? Which method is suitable for this calculation?

References

- Artiga Gonzalez, A., Bertschinger, R., Brosda, F., Dahmen, T., Thumm, P., & Saupe, D. (2019). Kinetic analysis of oxygen dynamics under a variable work rate. *Human Movement Science, 66*, 645–658.
- Clarke, D. C., & Skiba, P. F. (2013). Rationale and resources for teaching the mathematical modeling of athletic training and performance. *Advances in Physiology Education, 37*(2), 134–152.
- Fayazi, S. A., Wan, N., Lucich, S., Vahidi, A., & Mocko, G. (2013). Optimal pacing in a cycling time-trial considering cyclist's fatigue dynamics. In *American control conference* (pp. 6442–6447).
- Jones, A. M., & Poole, D. C. (2013). *Oxygen uptake kinetics in sport, exercise and medicine*. Routledge.
- Lipková, L., Kumstát, M., & Struhár, I. (2022). Determination of critical power using different possible approaches among endurance athletes: A review. *International Journal of Environmental Research and Public Health, 19*(13), 7589.
- Mongin, D., Chabert, C., Caparros, A. U., Guzmán, J. V., Hue, O., Alvero-Cruz, J. R., & Courvoisier, D. S. (2020). The complex relationship between effort and heart rate: A hint from dynamic analysis. *Physiological Measurement, 41*(10), 105003.
- Monod, H., & Scherrer, J. (1965). The work capacity of a synergic muscular group. *Ergonomics, 8*(3), 329–338.
- Morton, R. H. (1986). A three component model of human bioenergetics. *Journal of Mathematical Biology, 24*(4), 451–466.
- Morton, R. H. (1996). A 3-parameter critical power model. *Ergonomics, 39*(4), 611–619.
- Schulte, A., Kracht, P., Dörrscheidt, F., & Liesen, H. (1999). Modeling and simulation of the human exercise metabolism. In *Conference volume software and hardware engineering for the 21st century*, (pp. 377–382).
- Skiba, P. F., Jackman, S., Clarke, D., Vanhatalo, A., & Jones, A. M. (2014). Effect of work and recovery durations on W' reconstitution during intermittent exercise. *Medicine and Science in Sports and Exercise, 46*(7), 1433–1440.
- Sundström, D., & Bäckström, M. (2017). Optimization of pacing strategies for variable wind conditions in road cycling. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, 231*(3), 184–199.
- Wilkie, D. R. (1981). Equations describing power input by human as function of duration of exercise. *Exercise Bioenergetics and Gas Exchange, 75–80*.
- Wolf, S. (2019). *Applications of optimal control to road cycling*. Doctoral dissertation, University of Konstanz, Germany.
- Wolf, S., Biral, F., & Saupe, D. (2019). Adaptive feedback system for optimal pacing strategies in road cycling. *Sports Engineering, 22*(1), 1–10.
- Zignoli, A., Fornasiero, A., Bertolazzi, E., Pellegrini, B., Schena, F., Biral, F., & Laursen, P. B. (2019). State-of-the art concepts and future directions in modelling oxygen consumption and lactate concentration in cycling exercise. *Sport Sciences for Health, 15*(2), 295–310.
- Zignoli, A., Fornasiero, A., Ragni, M., Pellegrini, B., Schena, F., Biral, F., & Laursen, P. B. (2020). Estimating an individual's oxygen uptake during cycling exercise with a recurrent neural network trained from easy-to-obtain inputs: A pilot study. *PLoS One, 15*(3), e0229466.



Simulation of Physiological Adaptation Processes

Mark Pfeiffer and Stefan Ender

Contents

- 13.1 Example Sport – 100
- 13.2 Background – 101
- 13.3 Applications – 103
- References – 105

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Training and competition performance are strongly related to physiological adaptation processes.
- Higher level (macroscopic) models represent the overall effect of many individual processes (not exclusively physiological) on the body.
- The parameters of these metamodels can be individually adjusted (calibrated) based on measurable characteristics (data).
- Calibrated models allow short-term simulations (single training sessions or competitions) or long-term simulations (progress through training programs).

13.1 Example Sport

13

In various areas of sport, training is carried out to achieve goals in or through sport. If athletic training is geared toward athletic performance, the primary goal is to induce adaptation reactions that result in a physiological optimization of the functional systems involved. The accomplishment of sporting tasks in training represents a load acting on the organism, the extent of which results from the type and difficulty of the active task and the conditions of execution existing in this process. The load leads to a strain on the athlete, the individual extent of which depends on the use of the currently available physiological capacities (Hohmann et al., 2020). According to this stress-strain concept, a positive physiological adaptation takes place when the athlete exerts himself beyond a certain level (adaptive training stimulus). In this case, the fatigue (stress sequence) that initially occurs leads to an adaptation of the functional systems in the further course of recovery and thus to an optimization of performance. Whereas at a low level of performance, small increases in load or stress can bring about large increases in athletic performance or efficiency, at higher levels of performance, even greater stresses are necessary to still achieve increases in function or to maintain the current level. The particular

challenge in training now consists in triggering an adaptive training stimulus on the one hand and at the same time avoiding overloading of functional systems, because these can lead to maladaptations (Meeusen et al., 2013). The course of physiological adaptation processes in sports can be summarized as follows: physical stress leads to a strain on the loaded functional systems, whose capacities increase or are maintained in phases of recovery when the stimulus constellation is optimal. The scientific consideration of such processes can be directed to the relationship between load and strain during training (► Example 1) or the design of sequences of strain during successive training sessions, i.e., the relationship between load or strain and the altered performance (capacity) (► Example 2). In recent years, several informatics approaches have been introduced to model the relationship between load and stress or the adaptation reactions triggered by them and of investigating the effects of training strategies using simulation.

13.2 Background

Physiological adaptation as a consequence of athletic training is a complex, non-linear process. It has been shown that inter- and intraindividual variability is enormous, both in terms of physiological responses and effects on athletic performance or performance capacity (Borresen & Lambert, 2009). A repeated training stimulus can produce different effects in the same individual both during the loading situation and in the subsequent adaptation responses because the adaptation of a biological system itself leads to altered subsequent adaptations. Furthermore, physiological adaptation processes are highly dependent on a variety of other factors, such as human genetics, training status and history, psychological factors, and many others (Balagué et al., 2020; Pol et al., 2020). For the aforementioned reasons, the relationship between training (loading) and performance physiological responses (stress) both during (► Example 1) and in the aftermath of loading, as well as performance changes in the further course (► Example 2), is increasingly analyzed for the individual case using so-called meta-models. In their overview article, Rasche and Pfeiffer (2019) present different approaches for the analysis of individual time series in terms of the mathematical foundations and the underlying physiological assumptions in a comparative manner (■ Fig. 13.1).

From previous findings, assumptions on physiological adaptation processes are derived, which are represented to varying degrees in the model concepts (cf.

■ Fig. 13.1):

- athletic performance or capacity is a time-dependent state (output) that can be changed by training-induced physical stresses or strains (input)
- physiological adaptations occur with a time delay as a result of athletic training
- physical stress (training) affects the system state via two components: a negative one, which reduces the state variable output) and a positive one, which increases it (antagonism of the training effect) (Perl, 2002)
- physiological adaptations have an individual capacity limit
- negative and positive components are of identical magnitude

Approaches for Modelling the Relationship between Training (input) and Performance (output)						
Physiological Assumptions and Methods	<i>Time Dependency</i> I: Performance is a time-varying status II: Physiological adaptations / effects are delayed / decay over time					
	<i>Training Effects</i> III: Antagonistic effects on performance IV: Limited physiological capacities V: Equal positive / negative effects					
	Type B Difference Equation System (Flow Equation System)		Type A Differential Equation System (Exponential Decay)		Statistical Analysis Frequency-/time-domain Parametric /non-parametric Linear/non-linear	Machine learning/ Data mining: Classification and Clustering
Performance Models	Performance-Potential Double-Model	Performance-Potential Meta-Model	Impulse-Response Model + (Hill-Function) (also: Fitness-Fatigue /Banister-Model) Mod. Impulse-Response Model (+Kalman Filter) Mixed Linear Modelling		Auto-regressive / integrated /moving average models Auto- / Cross- Correlation	Artificial Neural Networks (Multilayer Perceptron)
	#Inputs	bivariate (2)	univariate (1)	univariate (1)	uni-/multivariate (1/n)	multivariate (n)

Fig. 13.1 Overview of selected models and their physiological assumptions, mathematical basis, and several input parameters. (After Rasche & Pfeiffer, 2019)

Summarizing the current state of the art in the informatics-oriented simulation of physiological adaptation processes, modeling approaches differ concerning their physiological assumptions and basic mathematical concepts. The empirical findings to date show that, depending on the quality (total number and frequency of input and output, quality of data collection methods, etc.) and the data collection setting (competitive or amateur sports), model adaptation to real data is quite satisfactory, but the quality of prediction is usually problematic. One of the reasons for the latter is that the model parameters determined during calibration are not “stable”, i.e. an almost identical goodness of fit (model fit) can be achieved with different parameter combinations (Hemingway et al., 2020). Thus, the model parameters can only be interpreted as an individual physiological fingerprint to a limited extent, which would, however, be important for the prognosis of future developmental courses or the investigation of training strategies (e.g., taper) (Vermeire et al., 2022).

Definition

Physiological simulation in sports is the calculation of the behavior of biological functional systems under physical stress (athletic training). The starting point for the simulation of physiological adaptation processes is a model according to which the trainee’s system has a current state that can be changed by a variable from outside (input). Both the system itself and its performance (output) can be described. The system behavior can be investigated cumulatively by varying the model parameters and data.

The simulation of physiological adaptation processes is used to

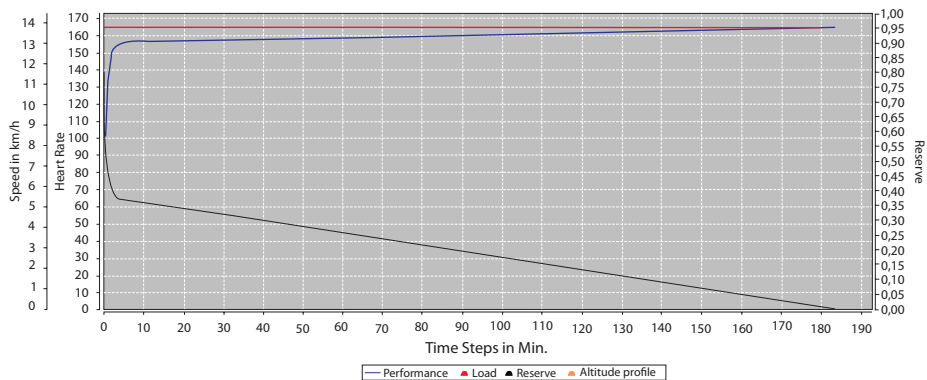
1. to better understand the trainee's behavior (including abnormalities) through parameter variation;
2. predict future performance developments;
3. Estimate the effects of virtual training strategies or scenarios.

13.3 Applications

► Example 1

The PerPot meta-model (Perl, 2001) was adapted by Endler (2013) specifically for endurance-oriented running (PerPot-run). On the one hand, this involved adapting the model itself by simplifying model equations that represent individual processes in the general metamodel that do not occur in this form in the application scenario of running sports (see ■ Fig. 13.2). On the other hand, properties important for endurance sports, such as long-term atrophy, i.e. fatigue over time, were integrated into PerPot-run. Based on the step test of the usual performance diagnostics, Endler (2013) developed and implemented a special calibration run to determine the individual model parameters. With the help of the individually adapted (calibrated) models, it is possible to perform simulations of training sessions and competitions. Thus, the optimal speed can be cumulatively determined by multiple simulations of a specific competition course with different (constant) speeds. The simulation is based on the fact that the athlete has virtually exhausted all his “energy” at the end of the competition. The target time calculated in this way was compared with the actual target times in a study of 33 marathons and half-marathons. The average deviation was 3.62%, whereby in more than one-third of the races the exact target time could be predicted. ◀

Another possibility to use the simulation already during training is the optimization of interval training units. The question often arises as to how many intervals



■ Fig. 13.2 Simulation of an optimal marathon run with constant speed

should be performed and at what speed. The length of the breaks between the intervals can also be chosen differently. With the help of simulations, these aspects can be examined and individual optimal interval training units can be identified.

► Example 2

In high-performance sports, the goal of training is to improve athletic performance, maintain it, or counteract a decline (e.g., due to age). To achieve this, it is necessary to repeatedly arrange training-effective stimulus constellations at the limit of the individual adaptation capacity, which causes performance-increasing or -maintaining adaptation reactions (adaptation-effective load). Targeted phases of recovery are necessary (Halson & Jeukendrup, 2004). If there is a permanent mismatch between training load and recovery, there is a risk of overtraining, i.e., athletic performance (ability) may stagnate or even temporarily decrease (non-functional overreaching); in the worst case, the stressed physiological functional systems only recover after months and performance permanently decrease (overtraining syndrome). This condition can cause pathological maladaptations in the various biological regulatory mechanisms and lead to an increase in susceptibility to injury or disease (Schwellnus et al., 2016; Soligard et al., 2016).

To design the training load and stress optimally in terms of training goal achievement and to avoid training-induced overloads, the individual training and performance data are examined for their correlation. These so-called training effect analyses are particularly helpful when the effects of training interventions on athletic performance (ability) can not only be modeled retrospectively but also simulated prognostically. This approach is particularly useful in sports or disciplines with less complex training and performance structures. Studies on cycling and swimming have shown that antagonistic models (cf. Fig. 13.1) can be used to model performance retrospectively (model fit) and to simulate future performance development satisfactorily for given training data, especially for shorter periods (forecast quality) (Pfeiffer & Hohmann, 2012; Pfeiffer, 2008; Fuhrmann et al., 2014). In recent years, a large number of model-comparative studies on model fit and prognostic goodness of fit have appeared, ranging from statistical approaches to machine-learning methods (including Imbach et al., 2022; Matabuena & Rodríguez-López, 2019). ◀

Study Box

The “PerPot-run” model presented in the “Areas of application” section in ► Example 1 can also be used to simulate the so-called individual anaerobic threshold (IAS). This is the load at which lactate build-up and breakdown in the body just balance each other out. To be able to reliably determine this value, athletes must complete endurance units at their subjective load limit over

several days. The load is increased slightly from unit to unit. Lactate is measured during the units. The threshold is then just the load at which the lactate concentration remains constant throughout the run and does not increase over time (lactate steady state). This method is not performed in practice due to the high demands placed on athletes. Instead, thresholds are deter-

mined using a step test as part of a performance diagnostic test. However, these tests are also time-consuming and cost-intensive, as medically trained personnel are required to perform them (blood sampling).

With PerPot-Run, a simulative determination of IAS was developed. In a study by Endler et al. (2017), this was compared with the most common performance diagnostic methods for determining IAS. For this purpose, 13 male handball players (age: 23.2 ± 2.3 ; weight in kg: 88.3 ± 11.4) completed a classic performance-diagnostic step test on a treadmill with a 3 min step length and a 2 km/h increase per step at a starting speed of 6 km/h. Between the steps, capillary blood was taken from the earlobe to determine lactate concentration. The

values were evaluated with different calculation models to determine IAS, including the most commonly used method by Dickhuth et al. (1999). Heart rate and speed data from the same step test were used to individualize the parameters of the PerPot-Run informatic model to the athletes. By simulating the previously described gold standard with several endurance runs, the IAS can then be calculated. A comparison showed a very high correlation between the heart rates determined with the Dickhuth method and the PerPot-Run at the IAS (ICC: 0.916; r : 0.889). It could thus be shown that the cumulatively determined thresholds represent a cost- and resource-saving alternative or supplement to classical lactate-based performance diagnostics.

? Questions for the Students

1. For what purpose can simulations of physiological processes be used?
2. What are antagonistic models and why are they used especially for the simulation of physiological adaptation processes?
3. Give examples of simulations of physiological adaptation processes that can be used in training.

References

- Balagué, N., Hristovski, R., Almarcha, M. D. C., Garcia-Retortillo, S., & Ivanov, P. C. (2020). Network physiology of exercise: Vision and perspectives. *Frontiers in Physiology*, *11*, 1607.
- Borresen, J., & Lambert, M. (2009). The quantification of training load, the training response and the effect on performance. *Sports Medicine (Auckland, N.Z.)*, *39*(9), 779–795.
- Dickhuth, H.-H., Yin, L., Niess, A., Rucker, K., Mayer, F., Heitkamp, H. C., & Horstmann, T. (1999). Ventilatory, lactate-derived and catecholamine thresholds during incremental treadmill running: Relationship and reproducibility. *International Journal of Sports Medicine*, *20*(2), 122–127.
- Endler, S. (2013). *Anpassung des Metamodells PerPot an den ausdauerorientierten Laufsport zur Trainings- und Wettkampfoptimierung*. Dissertation, Johannes Gutenberg-Universität Mainz. <https://doi.org/10.25358/openscience-3652>.

- Endler, S., Hoffmann, S., Sterzing, B., Simon, P., & Pfeiffer, M. (2017). The PerPot simulated anaerobic threshold: A comparison to typical lactate-based thresholds. *International journal of human movement and sports sciences*, 5(1), 9–15. <https://doi.org/10.13189/saj.2017.050102>
- Fuhrmann, S., Pfeiffer, M., & Hohmann, A. (2014). Modellierung von Trainingsprozessen im Schwimmsport. In M. Witt (Ed.), *DVS-Schwimmsport-Symposium 2011* (pp. 91–98). Deutsche Schwimmtrainer-Vereinigung.
- Halson, S. L., & Jeukendrup, A. E. (2004). Does overtraining exist? An analysis of overreaching and overtraining research. *Sports Medicine*, 34(14), 967–981.
- Hemingway, B., Burgess, K., Elyan, E., & Swinton, P. (2020). The effects of measurement error and testing frequency on the fitness-fatigue model applied to resistance training: A simulation approach. *International Journal of Sports Science & Coaching*, 15(1), 60–71.
- Hohmann, A., Lames, M., Letzelter, M., & Pfeiffer, M. (2020). *Einführung in die Trainingswissenschaft*. Limpert.
- Imbach, F., Perrey, S., Chailan, R., Meline, T., & Candau, R. (2022). Training load responses modelling and model generalisation in elite sports. *Scientific Reports*, 12(1), 1586.
- Matabuena, M., & Rodríguez-López, R. (2019). An improved version of the classical banister model to predict changes in physical condition. *Bulletin of Mathematical Biology*, 2019(81), 1867–1884.
- Meeusen, R., Duclos, M., Foster, C., Fry, A., Gleeson, M., Nieman, D., et al. (2013). Prevention, diagnosis and treatment of the overtraining syndrome: Joint consensus statement of the European College of Sport Science (ECSS) and the American College of Sports Medicine (ACSM). *European Journal of Sport Science*, 13(1), 1–24.
- Perl, J. (2001). PerPot: A metamodel for simulation of load performance interaction. *Electronic Journal of Sport Science*, 1(2).
- Perl, J. (2002). Adaptation, antagonism and system dynamics. In G. Ghent, D. Kluka, & D. Jones (Eds.), *Perspectives—The multidisciplinary series of physical education and sport science* (Vol. 4th ed, pp. 105–125). Meyer & Meyer Sport.
- Pfeiffer, M. (2008). Modeling the relationship between training and performance: A comparison of two antagonistic concepts. *International journal of computer science in sport*, 7(2), 13–32.
- Pfeiffer, M., & Hohmann, A. (2012). Applications of neural networks in training science. *Human Movement Science*, 31(2), 344–359.
- Pol, R., Balagué, N., Ric, A., Torrents, C., Kiely, J., & Hristovski, R. (2020). Training or synergizing? Complex systems principles change the understanding of sport processes. *Sports Medicine - Open*, 6(1), 28.
- Rasche, C., & Pfeiffer, M. (2019). Training. In A. Baca (Ed.), *Modelling and simulation in sport and exercise* (pp. 187–207). Routledge.
- Schwellnus, M., Soligard, T., Alonso, J. M., Bahr, R., Clarsen, B., Dijkstra, H. P., et al. (2016). How much is too much? (Part 2) International Olympic Committee consensus statement on load in sport and risk of illness. *British Journal of Sports Medicine*, 50(17), 1043–1052.
- Soligard, T., Schwellnus, M., Alonso, J. M., Bahr, R., Clarsen, B., Dijkstra, H. P., et al. (2016). How much is too much? (Part 1) International Olympic Committee consensus statement on load in sport and risk of injury. *British Journal of Sports Medicine*, 50(17), 1030–1041.
- Vermeire, K., Ghijs, M., Bourgois, J. G., & Boone, J. (2022). The Fitness–Fatigue model: What’s in the numbers? *International Journal of Sports Physiology and Performance*, 2022(17), 810–813.

Programming Languages

Contents

Chapter 14 An Introduction to the Programming
Language R for Beginners – 109

Robert Rein

Chapter 15 Python – 125

Maximilian Klemm



An Introduction to the Programming Language R for Beginners

Robert Rein

Contents

- 14.1 History and Philosophy – 110
- 14.2 Concept and Programming Paradigms – 111
- 14.3 Resources on R – 112
- 14.4 R Community and Packages – 112
- 14.5 Introduction to Working with R – 113
- 14.6 An Example Workflow in R – 116
- References – 123

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- The R programming language has been specifically designed for data analysis
- R allows an easy entry into data analytics without much previous programming experience
- R is ideal as a first programming language
- Through the use of packages, the functionality of R is constantly being extended so that a wide variety of data analyses can be easily performed
- Since R is open source, its usage is free of charge and future-proof
- R is available for all common operating systems and can be used largely independent of them
- With additional packages, R fully supports a literate programming approach and allows the easy creation of even complex reports in a wide variety of formats

14.1 History and Philosophy

14

The R programming language was developed with the aim of making data analysis as simple as possible. Historically, R originated as a free alternative to the S language (Ihaka, 1998). The S language was developed in the 1970s at Bell Laboratories, then part of AT&T. The developers of S wanted to create a language that makes data analysis as easy as possible. The conceptual approach was to develop a system that allows a gradual transition from user to developer. This was achieved by combining an interactive command line with facilities for creating classic programs (Venables & Ripley, 2000). The original S-System (later S-Plus) was, however, limited in its dissemination possibilities by its proprietary character. In 1991, therefore, Ross Ihaka and Robert Gentleman at the University of Auckland (New Zealand) began developing an alternative. The new programming language, named R, was first introduced in 1993 and licensed under the GNU General Public License in 1995. This made R a "free" software that allows further development away from

proprietary restrictions. In retrospect, this decision turned out to be the right one and is still one of the driving forces in the further development of R. Today, R is the franca lingua for statistical analyses in the field of university research and has largely replaced other proprietary statistical packages.

Another leap in the dissemination of R came with the release of RStudio (now Posit). RStudio is an integrated development environment based on R which simplifies the handling of various functionalities. RStudio is developed by RStudio PBC (public benefit corporation). The development environment has once again made access to R much easier for programming novices without, however, abandoning the core idea of interactive data analysis in combination with longer scripts or programs. Therefore, even with RStudio, data analysis using mouse-driven menus is not possible and users interact with R using commands, which is still the most efficient way of analyzing data.

Today, R runs on all standard operating systems (Linux, Windows, macOS) as well as on numerous non-standard systems. The further development of R is controlled by non-profit organization the R Foundation. The R Foundation is part of the Free Software Foundation and ensures that there are no restrictions due to proprietary hurdles and ensures the continuous further development of R. In the meantime, there are various commercial providers in the R space who provide support services for companies. In addition, many commercial statistical packages have implemented programming links to R.

14.2 Concept and Programming Paradigms

R is an interpreted programming language. This means that R executes programming commands immediately and the entire program does not have to be translated into machine commands by a compiler first, as is the case with the programming language C, for example. This has the advantage that the work can be done interactively, i.e. as a user, a result is returned by R after each command is executed. This simplifies data analysis to a great extent, as data can be quickly adjusted, transformed or displayed graphically or descriptively. The disadvantage of interpreted programming languages is that by processing the individual commands, certain compiler optimizations cannot be applied, so that the execution time can be longer compared to compiled programming languages. However, R can get around this disadvantage in many cases by including packages created in other programming languages.

Conceptually, R enables programming with different programming paradigms. One origin of R is derived from the language Scheme (Sussman & Steele, 1998). Scheme has the special feature that, somewhat simplified, there is no difference between data and code (Abelson & Sussman, 1996). Therefore, R allows for so-called meta-programming, i.e. programming on the language itself. This makes it relatively easy to create domain-specific sublanguages in R to simplify various tasks (Wickham, 2019). Over the years, numerous possibilities for classical object-oriented programming have also been added. The simple extensibility of R makes it possible to constantly develop the programming paradigms that can be used, so

that, for example, functional programming or newer paradigms such as reactive programming are also possible (Wickham, 2021). Despite the constant further development and expansion, R has always managed to keep the entry hurdle consistently low. Therefore, it is still possible to start a data analysis directly after a 2 min installation.

14.3 Resources on R

While it was difficult to get help with problems at the beginning of R, this has changed dramatically in the last 10–15 years. For example, due to the somewhat unfortunate naming of R for search engines, the search for solutions to problems was not very successful at the beginning. However, this has changed completely for the better. On the internet there are countless extremely active communities around R, with detailed blogs, podcasts, YouTube collections and programming help for specific questions (e.g. Stack Overflow). In addition, the number of books around and about R has exploded in recent years. With its own series on data analysis with R (Springer Use R!, CRC The R-Series) and countless other scientific books with at least code examples in R up to freely available collections of high-quality, scientific books on R (► bookdown.org). In addition, with the introduction of ChatGPT and similar LLMs it is really easy to obtain individual support.

Therefore, getting started in data analysis with R is easier than with any other programming language. In addition, within the scientific community in the field of statistics, R is the dominant language for implementing and testing new procedures. This leads to the fact that the latest statistical procedures are usually provided directly in R through packages developed by the respective scientists themselves and are thus available at an early stage.

14.4 R Community and Packages

14

A driving force in the further development and dissemination of R is the huge community of users and programmers. Because R is a complete programming language at its core, its functionality can be constantly expanded and adapted to individual needs. New functionality is bundled in R within the framework of so-called packages (alternatively libraries). Through these packages, new commands can be made accessible through newly defined functions in R.

R packages are distributed via the Comprehensive R Archive Network (CRAN for short). CRAN is an international network of web servers on which R packages are stored and which allow easy downloading from within R. All packages stored on CRAN adhere to a strictly defined structure and undergo quality control. Further development and adaptation is ensured by so-called maintainers. While the number of additional packages was still relatively manageable at the beginning of R, the current number of R packages on CRAN is 18,720 (as of 10.2022) with a constantly rising trend. Since data analyses across different disciplines and use cases are basically always similar, there is a high probability that existing packages

and additional functions in R are also available for unusual use cases. Therefore, for the majority of users, it is often no longer necessary to carry out complicated programming tasks themselves. Instead, by searching for a suitable package, problems that arise can be solved quickly. This also leads to the fact that the entry hurdle for dealing with R is very low.

14.5 Introduction to Working with R

As has been emphasised several times, dealing with R consists of having R execute specific commands. Thus, in the simplest case, R can be seen as an over-proportioned calculator. For example, on the R command line, the following command $2 + 2$ followed by an ENTER leads to the following sequence:

```
> 2 + 2
[1] 4
```

Here $>$ stands for the command line and $[1]$ indicates the first line of the output of R. The command line works according to the principle of a so-called REPL. REPL is an abbreviation for read-eval-print loop. The input is read in by R (R), evaluated within the programming language (E), the result is display (P) and subsequently the command line goes back to the initial state (L).

However, the 4 calculated in the example is now no longer available for further processing. Since R has executed the REPL and the output is not automatically saved. In order to further process the return value of an expression, this value must be made accessible in some form. In order to be able to further use calculated values, these values must be assigned an identifier (name). This entails the concept of a variable. Experience has shown that this concept represents a first major hurdle for the transition from, for example, spreadsheet programmes where the calculations seem to take place directly on the data to be seen. To assign a name to an expression or its return value in R, the assignment operator $<-$ is used. For example, if I want to give the result of the “complex” calculation $2 + 2 * 4$ a name, it would look like this:

```
> x <- 2 + 2 * 4
```

In this case, R does not return an expression, but has internally given the result of $2 + 2 * 4$ the identifier x . Calling x from on the command line then leads to:

```
> x
[1] 10
```

I.e. the value 10 is stored in the internal memory of R and the value can be called up or output via the identifier `x`. This is a fundamental difference in the way of working compared to spread sheet programmes. In R, calculations, the return values of expressions, are assigned identifiers and can then be called again in later steps. Conversely, if intermediate results do not have an identifier, they cannot be reused.

Two further explanations of the previous examples are necessary. In the previous expressions, spaces have been placed between the individual parts of the expressions. These spaces are only for readability and have no influence on the evaluation of the expression by R. Therefore, the expressions `2 + 2 * 4` and `2 + 2*4` are equivalent and lead to the same result. When outputting the value, you probably also noticed that R did not calculate the value 16, which would be correct if the evaluation of the expression is carried out strictly from left to right. However, R has applied the correct mathematical rule of multiplication before addition and has therefore arrived at the mathematical correct result of 10.

When further processing identifiers in R, note that R distinguishes between upper and lower case. Therefore, calling the identifier (upper case X):

```
> X
Error: object 'X' not found
```

Results in an error. The occurrence of errors often leads to great confusion for newcomers to R, but it is a completely normal occurrence in everyday programming and should therefore not upset anyone. In this case, R only complains that it cannot find the identifier X and therefore does not know how to proceed.

Working with R is largely based on the application of functions to values. In R, functions are used according to the pattern `<NAME>(<PAR1>,<PAR2>,...,<PARk>)` (the characters `<>` indicate any identifier). I.e. as soon as a pair of round brackets follows an identifier, R assumes that a function has to be called. Via `<PAR1>,<PAR2>,...,<PARk>`, comma-separated parameters can be passed to the function. The number of parameters depends on the definition of the function. A simple example is the application of the square root to a numerical value.

```
> y <- 9
> sqrt(y)
[1] 3
```

The mathematical square root function is realized in R by the `sqrt()` function. In the example, the identifier `y` is first assigned to the value 9 and the root function `sqrt()` is then applied to this identifier. An example somewhat closer to a real application would be, for example, the calculation of the mean value or the sum of the data series (3, 5, 7). In R, such an ordered series of numbers is represented as a

vector. To create such a vector, a function `c()` (c for concatenation) is used. Then the `mean()` or `sum()` function can be applied to the created vector.

```
> z <- c(3, 5, 7)
> z
[1] 3 5 7
> mean(z)
[1] 5
> sum(z)
[1] 15
```

The great advantage of R is that you can easily define your own functions. For example, a function that returns the minimum and maximum of a vector as a vector with two entries can be defined as follows.

```
> my_min_max <- function(x) {
+   c(min(x), max(x))
+ }
> my_min_max(z)
[1] 3 7
```

Here the keyword `function()` R indicates that a new function is being defined. The keyword is followed by the two round brackets with the required parameters. In the example, only one parameter is required, which is given the identifier `x`. The naming is completely arbitrary and only has to be used if the function is to be defined. The naming is completely arbitrary and must only be used appropriately in the following function body, which is defined by the curly brackets `{}` area. When the function is called, the parameter is replaced according to the value passed in the brackets to the function body. The `+` in the output shown are just formatting symbols by the R command line and should not be typed in by the programmer.

If a function from an R package is required. The package must first be installed in the R environment on the local computer, if this has not already been done in a previous session. Again, a function is used for this purpose. For example, to generate interactive maps with R, the package `leaflet` is necessary. The following command installs the package in the R system:

```
> install.packages("leaflet")
```

R contacts the CRAN server in the background and downloads the corresponding package and required dependencies. The functionality of the package is then not yet directly available, but the package must first be loaded into the currently active working environment with another command.

```
> library(leaflet)
```

These examples only serve to give a very first overview of working with R and to get to know the first concepts of working with R. As can be seen from this simple example, one of the challenges is to learn the necessary commands and functions. Getting started with R is therefore similar to learning a new (very simple) language. Nowadays, this initial hurdle is made much easier by the significantly improved search functions of internet search engines. Thus, a search that starts with an R and the problem is usually sufficient to find suggested solutions. In this way, rapid productivity in R can be achieved without the need for in-depth programming knowledge. Therefore, R is very well suited as a first programming language in which more advanced programming concepts can be gradually developed as needed. Good starting sources for getting started are (Chambers, 2008; Dalgaard, 2020; Peng, 2016; Wickham & Grolemund, 2016).

The following is a somewhat more extensive example with only brief explanations. Detailed explanations of the commands used can be obtained in R through the help documentation. To do this, simply place a ? in front of the function name and R opens the corresponding help file.

```
> ?mean
```

14.6 An Example Workflow in R

Let the following data set from [Table 14.1](#) be given. In two independent groups A and B, the body fat content was determined and now it is to be examined whether there is a statistically significant difference between the two groups. Of course, this is only a synthetic example and should therefore not be carried out in this form within an actual scientific paper, but only serves as an illustration ([Table 14.1](#)).

In order to carry out a data analysis, the data must first be loaded into R. In the raw form (Supplementary Material: `bfp_data.txt`), the data are in the form of a text file. The first column of the file shows the group membership, while the second column contains the respective fat content. The columns are separated by a comma and a dot is used as decimal separator, following international conventions. To load the data into R, the `read_csv()` function from the `readr` package is used. First, however, the current working directory of R must be set to the corresponding folder under which the file is stored. The function `setwd()` (short for set working directory) is used for this. Since path specifications in R are always determined relative to the working directory, this step makes further work easier since no long file paths have to be specified.

```
> setwd(<PFAD>)
```

■ **Table 14.1** Percent body fat content in two groups

A	B
13.3	22.0
6.0	16.0
20.0	21.7
8.0	210.0
14.0	30.0
19.0	26.0
	30.0

Now the package `readr` is loaded and the file is loaded using the function `read_csv`.

```
> library(readr)
> bfp <- read_csv(file = 'bfp_data.txt')
```

The data is now available in R under the identifier `bfp`. The name is chosen arbitrarily and kept short for the sake of simplicity.

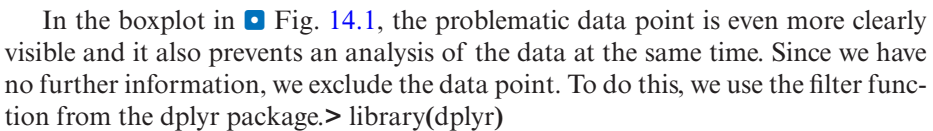
```
> bfp <- read_csv(file = 'bfp_data.txt')
# A tibble: 13 x 2
  Group   BFP
  <chr> <dbl>
1 A     13.3
2 A      6
3 A    20
4 A      8
5 A    14
6 A    19
7 B    22
8 B    16
9 B   21.7
10 B   210
11 B    30
12 B    26
13 B    30
```

The data is stored in a so-called data.frame object (or the newer version tibble) and can now be processed further. For example, an overview of descriptive statistics of the data can be generated using the function `summary()`.

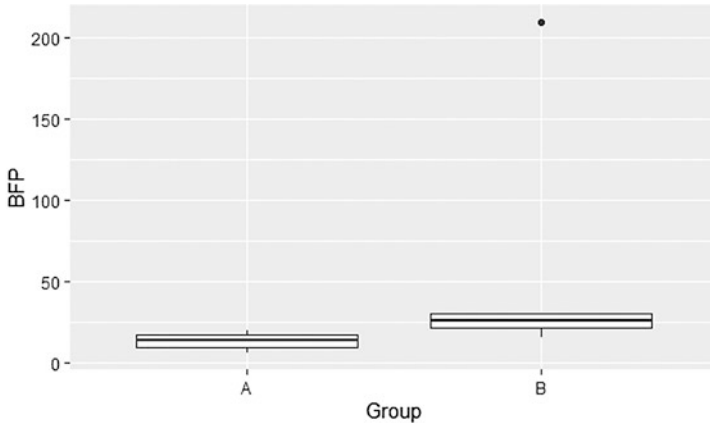
```
> summary(bfp)
  Group          BFP
Length:13      Min.   :  6.00
Class :character 1st Qu.: 14.00
Mode  :character Median : 20.00
                Mean  : 33.54
                3rd Qu.: 26.00
                Max.  :210.00
```

Here you can already see that one of the data points is probably incorrect because the value is >100 , which is not possible for a percentage body fat. In the next step, the data is displayed graphically using a boxplot. R provides numerous functions for simple graphical representation. However, we will use the package `ggplot2` here, which enables the creation of modern publication-quality graphs (Healy, 2018; Wickham, 2016). Again, the package must first be loaded before its functionality can be accessed.

```
> library(ggplot2)
> ggplot(bfp, aes(Group, BFP)) + geom_boxplot()
```

In the boxplot in  Fig. 14.1, the problematic data point is even more clearly visible and it also prevents an analysis of the data at the same time. Since we have no further information, we exclude the data point. To do this, we use the filter function from the `dplyr` package. `> library(dplyr)`

```
> bfp_clean <- filter(bfp, BFP <= 100)
> bfp_clean
# A tibble: 12 x 2
  Group   BFP
<chr> <dbl>
1 A      13.3
2 A         6
3 A      20
4 A         8
5 A      14
6 A      19
7 B      22
8 B      16
9 B     21.7
10 B      30
```



■ Fig. 14.1 Illustration of the sample data using a boxplot with the problematic data point

```
11 B      26
12 B      30
> ggplot(bfp_clean, aes(Group, BFP)) + geom_boxplot()
```

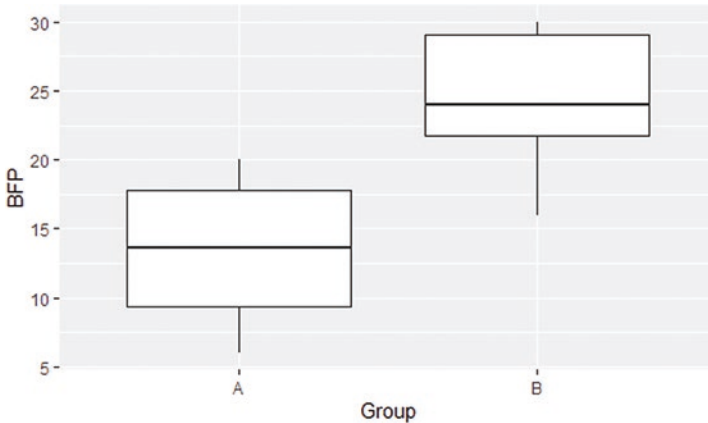
The boxplot is now much more informative (see ■ Fig. 14.2). Without going into the statistical prerequisites any further, we now conduct an independent t-test for groups with different variances. For this we again use a function from R.

```
> t.test(BFP ~ Group, data = bfp_clean)

Welch Two Sample t-test

data:  BFP by Group
t = -3.4017, df = 9.9886, p-value = 0.006762
alternative hypothesis: true difference in means between
group A and group B is not equal to 0
95 percent confidence interval:
 -18.040619  -3.759381
sample estimates:
mean in group A mean in group B
    13.38333      24.28333
```

As this example shows, a data analysis can be realised in R with just a few commands. The developers of R have taken care that the naming of functions is as close as possible to the desired activity, so that the English term usually allows one to quickly deduce the function.



■ Fig. 14.2 Representation of the sample data by means of a boxplot after exclusion of the outlier

In the example, we entered all commands directly on the command line and analysed the data interactively. In an actual analysis, the data analysis will consist of a combination of interactive work and permanent scripts. For example, those final commands to be applied to the data would be written to a script file so that the analysis can be revisited or retraced at a later time. In this way, the workflow shown could lead to the following script:

```
setwd(<PFAD>)

# Required packages
library(readr)
library(ggplot2)
library(dplyr)

# Read in data
bfp <- read_csv(file = 'bfp_data.txt')

# Process data
bfp_clean <- filter(bfp, BFP <= 100)

# Descriptive analysis
summary(bfp_clean)

# Graphics
ggplot(bfp_clean, aes(Group, BFP)) + geom_boxplot()

# Analysis
t.test(BFP~Group, data = bfp_clean)
```

The commands used are even more comprehensible through the use of comments, which are signaled in R with a #.

Since the traceability of data analysis not only for oneself but also by external parties in the context of scientific reproducibility is becoming increasingly important in scientific practice, R offers extensive functionality for documenting and publishing data analyses (Xie, 2016, 2017). Conceptually, this area falls under so-called literate programming. In 2002, a special package Sweave was published which allows the combination of R and Latex code in a single document. Building on this, a whole standard has since developed in the form of RMarkdown documents to generate reports, scientific articles, books, web blogs, dashboards, interactive documents, presentations and much more directly from R. A wide variety of output formats (e.g., text, graphics) are available for this purpose including for example docx, pdf, odf, pptx, html.

At its core, many of these processes are based on the Markdown standard, which makes it possible to easily create even the most complex documents. Markdown is a simplified markup language that can be used to create documents without complicated word processing systems and where the final formatting of the documents is left to an external systems (the pandoc software in R). The RMarkdown format enriches the Markdown format with R specific elements. A simple example document based on our previous scripts would be:

```
---
author: "Robert Rein"
title: "RMarkdwon-Beispiel"
output: pdf_document
---

# Required packages
```{r}
library(readr)
library(ggplot2)
library(dplyr)
```

# Read in data
```{r}
bfp <- read_csv(file = 'bfp_data.txt')
```

# Process data
```{r}
bfp_clean <- filter(bfp, BFP <= 100)
```
```

```

# Deskriptive analysis
```{r}
summary(bfp_clean)
```

# Graphics
```{r}
#| fig.cap="Boxplot of data",
#| fig.height=3
ggplot(bfp_clean, aes(Group, BFP)) + geom_boxplot()
```

# Analysis
```{r}
t.test(BFP~Group, data = bfp_clean)
```

```

The RMarkdown document begins with a so-called YAML header with meta data. This is followed by the actual document. In Markdown notation, a hash (#) means a first-order heading. The areas between `{r}` and ````` denote an R-code area called a chunk. The document can be translated into a PDF report using the knitr package. R first executes the respective code areas and the corresponding expressions and values are replaced in the chunks. Generated graphics are also inserted directly into the document. Recently, an updated version of RMarkdown called Quarto was introduced which further improved the combination of code and text and further allows to combine different programming languages in a single document.

Overall, the R environment therefore allows the complete life cycle of a data analysis from the initial processing to the publication of the data from a single environment. This allows recurring workflows to be automated and reporting systems to be created relatively easily for a wide range of use cases.

Study Box

Data analysis is rarely a straightforward task in which all processing steps are fixed from the beginning. Therefore, every data analysis is characterized by interactive phases in which the data must first be examined and processed. Following data processing, the actual data analysis usually takes place. These two steps can be run through several times in iterative cycles before a final process pipeline is created with which the data are finally analyzed and published. A software environment for data analysis must therefore optimally support these steps. The R programming language was designed precisely with the aim of enabling these phases and therefore offers possibilities for interactive as well as more structured ways of working. Through the possibility of integrating additional packages, the functionality of R is constantly

being expanded and for practically all possible use cases there are, if not complete, then already ready-made partial solutions. R offers extensive support for the complete life cycle of a data analysis. Since R is a complete programming language, all analysis steps can be adapted to one's own needs. R allows application at different levels of complexity as needed. Thus, in the course of time, a development from a new user with no programming experience to a fully-fledged programmer can be undergone.

? Questions for the Students

1. what distinguishes working with R from spreadsheet programs?
2. How can the functionality of R be easily extended?
3. what is meant by the term literate programming?
4. what happens with the following input?

```
> y <- 3
> z <- 3 * y + 4
> z
```

References

- Abelson, H., & Sussman, G. J. (1996). *Structure and interpretation of computer programs*. The MIT Press.
- Chambers, J. M. (2008). *Software for data analysis: Programming with R* (Vol. 2). Springer.
- Dalgaard, P. (2020). *Introductory statistics with R* (2nd ed.). Springer.
- Healy, K. (2018). *Data visualization: A practical introduction*. Princeton University Press.
- Ihaka, R. (1998). R: Past and future history. *Computing Science and Statistics*, 30, 392–396.
- Peng, R. D. (2016). *R programming for data science*. Leanpub.
- Sussman, G. J., & Steele, G. L. (1998). Scheme: A interpreter for extended lambda calculus. *Higher-Order and Symbolic Computation*, 11(4), 405–439.
- Venables, W., & Ripley, B. D. (2000). *S programming*. Springer Science & Business Media.
- Wickham, H. (2016). *Programming with ggplot2*. Springer.
- Wickham, H. (2019). *Advanced R*. CRC Press.
- Wickham, H. (2021). *Mastering shiny*. O'Reilly Media, Inc.
- Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc.
- Xie, Y. (2016). *Bookdown: Authoring books and technical documents with R markdown*. Chapman and Hall/CRC.
- Xie, Y. (2017). *Dynamic documents with R and knitr*. Chapman and Hall/CRC.



Python

Maximilian Klemp

Contents

15.1 Example Sport – 126

15.2 Background – 127

15.3 Applications – 129

References – 130

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Python is a dynamic, object-oriented, high-level open-source programming language and one of the most popular languages for Data Science tasks
- The large volume of data generated in sports require efficient and flexible analysis pipelines, spanning from large-scale data pre-processing (such as position and event data) to modelling by means of Machine Learning or Artificial Intelligence
- Python offers a wide range of functionalities through its various libraries, providing solutions for every processing step occurring in sports analytics
- Multiple state-of-the-art frameworks frequently used in sports analytics are implemented in Python and accessible for researchers, teachers and students via open source libraries
- the numerous use cases show the importance and significance of Python in sports analytics

15.1 Example Sport

15

The relevance of programming languages in the sports sciences is mainly rooted in the digital revolution in sport, which has been initiated by the development of efficient data collection methods for various data types during the past few decades (Memmert & Raabe, 2018). A football match, for instance, generates roughly seven million data points of position data and event data on around 1800 events, while depending on the data provider these events are described in up to 20 dimensions. When these numbers are multiplied by the number of matches which can potentially be analysed (e.g. 380 matches per season in most European football leagues), the need for efficient data processing quickly becomes apparent. In particular, processing and analysis tasks consist of filtering and aggregating data points, calculat-

ing or transforming variables as well as the subsequent modelling of relationships (for more detail, refer to ► Sect. 15.2). In this regard, recurring analysis steps are needed, which for example have to be applied to newly generated data or, due to the total volume of the analysed data, have to be performed repeatedly on subsets of the data. For this reason, the development of robust and reproducible data processing pipelines via ad-hoc scripts is inevitable. Among Data Scientists, Python enjoys great popularity as the programming language of choice owing to its features, which are discussed in the next section. Also for the various use cases in sports analytics, Python might have come out on top as a programming language.

15.2 Background

From the previously discussed structure and volume of data, there arise specific requirements for the data processing tasks in sports analytics. For example, oftentimes raw data (e.g. position or event data) are available for multiple matches, training sessions or sequences, which have to be pre-processed before actual analysis steps can be performed. This pre-processing might include filtering, grouping and aggregation operations as well as creating variables by transforming or combining existing variables. Sometimes it is necessary to join data from different sources on a common index. For example, position and event data might have to be synchronized on the level of a timecode or meta-information of matches might have to be combined with variables calculated from raw data (following the terminology from the database querying language SQL, these operations are referred to as “joins”). This bandwidth of possible processing steps already makes the development of efficient and robust pre-processing pipelines inevitable. Furthermore, in processing and analysing sports data, often the same steps have to be performed repeatedly. The reason for this might be the collection of new data, which has to be processed immediately, or the vast volume of the analysed data. Position data for one football match takes around 500 MB of memory in Python, so loading a whole season (which amounts to a sample size that analyses should be based upon at best) at once is impossible. This circumstance primarily justifies the need for reproducible analysis routines. Finally, after creating structured datasets through the steps outlined above, this data has to be modelled using methods of Machine Learning or Artificial Intelligence in order to find associations between indicators or to solve classification problems in the domain of computer science in sports (such as the generation of position data from broadcasting videos or the detection of match phases).

Python offers solutions for all of the challenges named above and is, due to this and other features, well-suited for data processing and analysis in the sport context. Python is highly popular in the community of data scientists in general and sports analysts in particular because it is a open-source, dynamic, object-oriented, high-level programming language, which provides highly flexible and up-to-date functionalities due to its available modules and libraries. With respect to the

above-mentioned necessities, the following libraries are particularly valuable: *pandas* (McKinney, 2011) provides functionalities for data import from different sources and illustrative presentation. *NumPy* (Oliphant, 2006) enables efficient calculations, including numerous vector and matrix operations, whereas *SciPy* (McKinney, 2010) provides functions for calculus and time series analysis. As an extension to the rudimentary visualizations in *Pandas*, the package *Matplotlib* (Barrett et al., 2005) contains a variety of possibilities for data visualization. Finally, via *Statsmodels* (Seabold & Perktold, 2010) one can employ “traditional” statistical models such as linear models and generalized linear models, while in *scikit-learn* (Pedregosa et al., 2011) algorithms from Machine Learning can be implemented. Having all those functionalities available simultaneously along with the dynamic nature of Python, enabling interactive programming and therefore quick and flexible analyses, makes the language especially valuable for data analysis and processing tasks. The high level of abstraction furthermore makes learning the language from scratch easier.

Definition

Python is a *dynamic, object-oriented, high-level, open-source* programming language. *Dynamic* refers to the compiling process. Compiling a program includes the translation from source code (written in the respective programming language) to machine-readable binary code. Every programming language has to be compiled in this way, whereas for dynamic languages this step does not have to be explicitly performed by the developer but is done line-by-line by the so-called *Interpreter*. The consequence of this is that in Python commands can be executed independently from the rest of the program, which enables interactive programming and therefore immediate testing of code snippets and quick data analyses. *Object-oriented* programming languages work with so-called classes, which define features and functions in a generalized way and therefore allow reproducible analyses of imported data. *High-level* programming languages are a stronger abstraction from the specifics of one particular machine. The language syntax has a better readability and interpretability for humans, which makes Python more user-friendly than other languages. The downside of dynamic and high-level languages is a reduced efficiency, which manifests in a slower processing speed. A way to address this issue in Python is to embed modules, which are implemented in other, more efficient programming languages such as C++ or C. Lastly, Python is *open-source* and therefore freely available and due to the community-based maintenance always held up-to-date. Another consequence of this is the presence of comprehensive internet resources for troubleshooting and tutoring, which greatly simplifies programming in Python.

15.3 Applications

► Example 1

The above-mentioned pre-processing steps (filtering, grouping, aggregation) are needed in multiple endeavours of analysing large data volumes. Accordingly, Anzer and Bauer (2021) created an Expected Goals Model using information from position and event data simultaneously. To this end, complex processing steps for synchronization of both data types were necessary. Klemp et al. (2021) collected multiple performance indicators separately for both halves of football matches and filtered variables with respect to ball possession or running velocities. Subsequently, the indicators calculated from raw position data had to be combined with those from raw event data and from another data source for betting odds. The operations performed in these two research articles consisted of several pre-processing steps, which were performed using the functionalities of *pandas* and *NumPy*. ◀

► Example 2

Another prominent use case within the field of tactical analysis comprises the calculation of the so-called variables of collective behaviour. These variables model the relationships and interactions among players and teams as geometric or algebraic entities, making the library *SciPy* a popular choice for these calculations. Examples of this can be found in the comparison of team formations by Memmert et al. (2019) or in the examination of substitutions in football by Lorenzo-Martínez et al. (2022). ◀

► Example 3

Finally, Python enables statistical modelling both in the sense of null hypothesis significance testing (e.g. Bassek et al., 2022) and for implementing complex models from the area of Machine Learning. One of the probably most well-known applications can be found in the work of Decroos et al. (2019), who built a model using event data to quantify the game state for any given time instance. The game state thereby comprises the probability of a team scoring or conceding a goal within a defined time interval in the future. Building upon this, all player actions can be valued based on their effect on the game state. The approach also contains data processing and visualization techniques. It is to be highlighted especially because the algorithms developed in the course of the study have been published within the Python library *socceraction* (► <https://github.com/ML-KULeuven/socceraction>) so that interested programmers can reproduce the analyses using data from different providers. Finally, in this respect, the publication of a large scale data set of event data by Luca Pappalardo et al. (2019) should be mentioned, where also functions for the import and analysis of the data are provided by the authors (► <https://github.com/Friends-of-Tracking-Data-FoTD/mapping-match-events-in-Python>). ◀

Study Box

As pointed out during the previous sections, data analysts in sports are confronted with recurring tasks, which in specific cases might still require customized solutions. To account for this circumstance and provide a variety of generalized solutions, the Python library *floodlight* was published by Raabe et al. (2022). It contains different pre-processing and analysis routines for the investigation of sports data, at this point in time specifically football and handball data. The functionalities of *floodlight* contain import functions for different data sources of position, event and meta data, visualization methods as well as specific data models from different disciplines like exercise physiology (Metabolic Power Model, di Prampero & Osgnach, 2018), dynamical systems (Approximate Entropy, Pincus, 1991) or collective behaviour (Bourbousson et al., 2010). The aim of *floodlight* was to standardize or externalize recurring processes of import and pre-processing so researchers could focus more specifically on the analysis itself.

? Questions for the Students

1. Why is Python so popular for Data Science tasks?
2. Name a specific use case for the libraries *pandas* and *scikit-Learn* from the realm of data analysis in sport.

References

- Anzer, G., & Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*, 3, 624475.
- Barrett, P., Hunter, J., Miller, J. T., Hsu, J.-C., & Greenfield, P. (2005). matplotlib—A portable python plotting package. In *Astronomical data analysis software and systems XIV*.
- Bassek, M., Raabe, D., Memmert, D., & Rein, R. (2022). Analysing motion characteristics and metabolic power in elite male handball players. *Journal of Sports Science and Medicine*, 22(2), 310–316.
- Bourbousson, J., Sève, C., & McGarry, T. (2010). Space–time coordination dynamics in basketball: Part 2. The interaction between the two teams. *Journal of Sports Sciences*, 28(3), 349–358.
- Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*.
- Di Prampero, P. E., & Osgnach, C. (2018). Metabolic power in team sports—Part 1: An update. *International Journal of Sports Medicine*, 39(08), 581–587.
- Klemp, M., Wunderlich, F., & Memmert, D. (2021). In-play forecasting in football using event and positional data. *Scientific Reports*, 11(1), 1–10.
- Lorenzo-Martínez, M., Rein, R., Garnica-Caparrós, M., Memmert, D., & Rey, E. (2022). The effect of substitutions on team tactical behavior in professional soccer. *Research Quarterly for Exercise and Sport*, 93(2), 301–309.
- McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference*.
- McKinney, W. (2011). pandas: A foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14(9), 1–9.

- Memmert, D., & Raabe, D. (2018). *Data analytics in football: Positional data collection, modelling and analysis*. Routledge.
- Memmert, D., Raabe, D., Schwab, S., & Rein, R. (2019). A tactical comparison of the 4-2-3-1 and 3-5-2 formation in soccer: A theory-oriented, experimental approach based on positional data in an 11 vs. 11 game set-up. *PLoS One*, *14*(1), e0210191.
- Oliphant, T. E. (2006). *A guide to NumPy* (Vol. Vol. 1). Trelgol Publishing USA.
- Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, *6*(1), 1–15.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.
- Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, *88*(6), 2297–2301.
- Raabe, D., Biermann, H., Bassek, M., Wohlan, M., Komitova, R., Rein, R., Groot, T. K., & Memmert, D. (2022). floodlight—A high-level, data-driven sports analytics framework. *Journal of Open Source Software*, *7*(76), 4588.
- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th python in science conference*.

Data Analysis

Contents

- Chapter 16** **Logistic Regression – 135**
Ashwin Phatak
- Chapter 17** **Time Series Data Mining – 141**
Rumena Komitova and Daniel Memmert
- Chapter 18** **Process Mining – 149**
Marc Garnica Caparrós
- Chapter 19** **Networks Centrality – 157**
*João Paulo Ramos, Rui Jorge Lopes,
Duarte Araújo, and Pedro Passos*
- Chapter 20** **Artificial Neural Networks – 169**
Markus Tilp
- Chapter 21** **Deep Neural Networks – 177**
Dominik Raabe
- Chapter 22** **Convolutional Neural Networks – 185**
Yannick Rudolph and Ulf Brefeld
- Chapter 23** **Transfer Learning – 193**
Henrik Biermann
- Chapter 24** **Random Forest – 201**
Justus Schlenger

- Chapter 25** **Statistical Learning for the
Modeling of Soccer Matches – 209**
Gunther Schauberger and Andreas Groll
- Chapter 26** **Open-Set Recognition – 217**
Ricardo da Silva Torres



Logistic Regression

Ashwin Phatak

Contents

16.1 Example Sport – 136

16.2 Background – 137

16.3 Application – 138

References – 140

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Logistic regression is one of the basic statistical classification models used for modeling binary outcomes.
- With a large amount of notational/event data available in sports, Logistic regression (LR) is one of the simplest tools to investigate a wide array of binary problems relevant in sports.
- Logistic regression can handle data imbalances, it is highly interpretable and computationally inexpensive.

16.1 Example Sport

In the sports industry, there are a wide variety of situations where modeling game event is a binary outcome (Mattera, 2021). Predicting win or loss probability in a sporting event, Goal or no goal, and foul or no foul are some scenarios in invasion sports where LR can be used. In individual sports such as tennis or badminton if the ball/shuttle was in or out. Such problems can potentially be modeled by a binary logistic regression, which is one of the simplest and most interpretable classification algorithms out there. Across different sports, binary classification modeling has been successfully used in order to investigate recruiting, coaching, self/

opponent analysis, and injury prediction based on ‘Big Data’ collected in the respective sports (Phatak et al., 2021).

16.2 Background


The rise of ‘Big Data’ in sports performance analysis has taken a whole new approach. Data collection mechanisms and their delivery to the sports industry have given rise to a wide potential for use of available data across sports and their respective sub-domains (Rein & Memmert, 2016). Statistical analysis and machine learning are the tools which help experts mine and interpret data in the decision-making and knowledge-discovery process (Phatak et al., 2021).

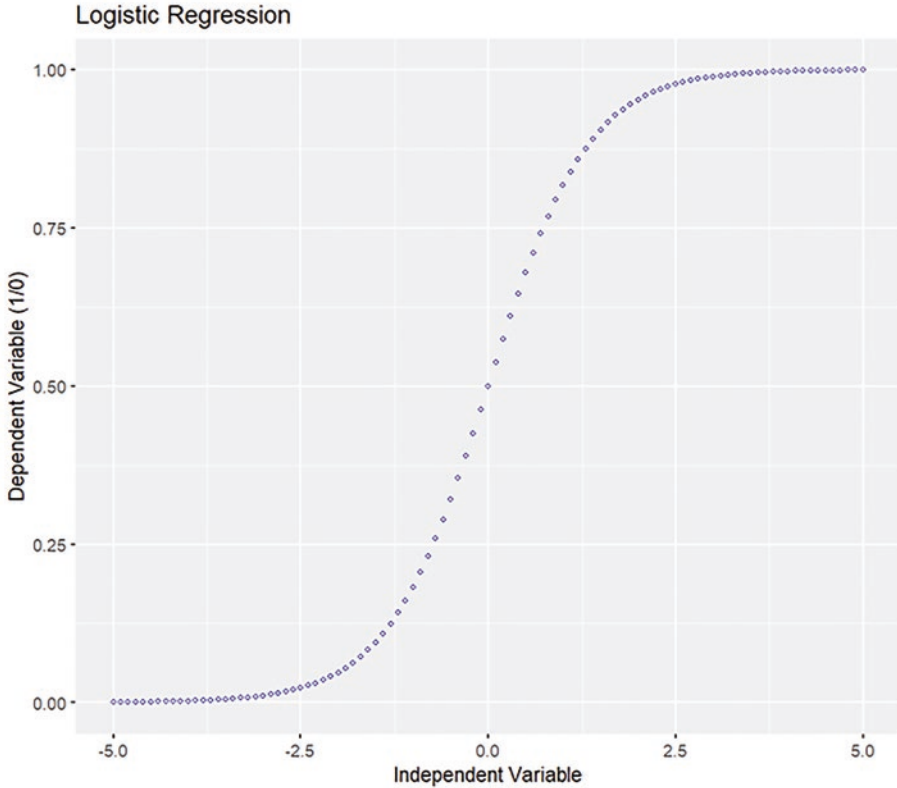
Classification and regression methods are the primary methods used for knowledge discovery. In classification, Logistic Regression is one of the simplest algorithms used for modeling binary problems. It is computationally inexpensive, interpretable, and robust against data imbalances. Programs like SPSS and Excel and languages like python and R have built-in libraries to implement Logistic Regression, making it simple and accessible to data analysts with different levels of technical skill (Persson, 2022).

Definition

In statistics, the logistic model is a statistical model that models the probability of an event. This is performed by taking the log odds for the event to be a linear combination of one or more independent variables (Wright, 1995). In essence, it is a linear regression compressed between 1 and 0 using a logit transform (see Equation below). So, the ceiling value is fixed at 1 and the floor (bottommost) at 0.

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \dots + \beta_n x_n)}}$$

Visually it can be interpreted as fitting a squiggle (see  Fig. 16.1) to the given data to predict a binary dependent variable. Logistic regression and its variations such as multiple logistic regression and logistic mixed models are capable of taking in multiple independent variables as input and also capable of modeling interaction effects.



■ Fig. 16.1 Logistic regression curve in two dimensions

16.3 Application

► Example 1

Let's consider a situation in an invasion game like football (soccer). There has been great debate about Video Automated referees (VAR) (Tovar, 2021). We can potentially model an automated system like VAR using Logistic Regression. In that case, the parameters will be as follows:

- Dependent variable: Offside = 1 and onside = 0
- Independent Variables: At the moment when the pass is played:
 - Vertical Location of the furthestmost body part of the second-last player of the defending team at the moment when the pass was played.
 - Vertical Location of the furthestmost body part of the highest player (involved in the game) of the attacking team at the moment when the pass was played.

We can train the logistic regression algorithm by giving it a label based on whether the given situation is onside or offside. We can then use the model as a rudimentary VAR system to detect offsides. ◀

► Example 2

Win/Loss probability can also be modeled using Logistic Regression in multiple sports based on betting odds (Wunderlich & Memmert, 2018; Wunderlich & Memmert, 2016). A model can be made to improve betting odds based on the results of the past 3 games. Take any sport in which team A is playing team B following would be the input and output parameters:

- Dependent Variable: Win = 1, Loss = 0
- Independent Variables:
 - Betting odds of team A winning
 - Betting odds of team B winning
 - Result of the last 3 games for team A
 - Result of the last 3 games for team B

Note: Such a model can be theoretically made but there are nuances. In our case, we have assumed that betting odds don't already account for the last three games. If this is the case we would have to deal with multicollinearity in the Dependant Variables which is a whole new topic by itself. You can look into it further if you are interested. ◀

► Example 3

Every season in the NBA, there is a draft of players from the NCAA. A model using logistic regression can potentially be designed to decide whether to pick the player or not (Liu et al., 2018). Following would be the parameters of such a model.

- Dependent Variable: Pick = 1, Not to Pick = 0
- Independent Variables:
 - Attacking Statistics
 - Field Goals
 - 3-point shots
 - 2-point shots
 - etc.
 - Defending Statistics
 - Rebounds
 - Steals
 - Blocks
 - etc.
 - Physiological Statistics
 - Height
 - Jump height
 - etc. ◀

Study Box

A study conducted on elite and sub-elite goalkeepers in soccer, analyzed, what set of Key Performance Indices (KPIs) distinguish champion league-level GKs (CL) from non-champion League (NCL) level ones (Jamil et al., 2021). The analysis involved the use of logistic regression where the CL GKs were encoded as 1 while the NCL

were encoded as 0 as the binary dependent variable. After some preprocessing steps, a set of 20 GK performance statistics (Independent Variables) were used to model the differences. It was observed that CL GKs were better at short distribution with their feet as compared to the NCL GKs. There seem to be no differences in the shot-stopping ability.

This Idea can be used to analyze any position in football (soccer) and the encoding of the players as 1/0 (success criteria) is also dependent on the research question. A question such as what is the difference between the KPIs of Relegation level midfielders as opposed to non-relegation level midfielders can also be answered using the same data and logistic regression by simply changing the encoding criteria of the players.

Questions for the Students

1. What are the maximum and minimum values can the dependent variables take in a Logistic Regression?
2. Give one example from three different sports where Logistic Regression can be used to model a specific scenario in respective sports. Note: List Dependent Variables and independent Variables for each scenario.

References

- Jamil, M., Phatak, A., Mehta, S., Beato, M., Memmert, D., & Connor, M. (2021). Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men's football. *Scientific Reports*, *11*(1), 1–7.
- Liu, Y., Schulte, O., & Li, C. (2018). Model trees for identifying exceptional players in the NHL and NBA drafts. In *International workshop on machine learning and data mining for sports analytics* (pp. 93–105). Springer.
- Mattera, R. (2021). Forecasting binary outcomes in soccer. *Annals of Operations Research*, 1–20.
- Persson, I. (2022). Review of applied univariate, bivariate, and multivariate statistics using python by Daniel Denis. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*, 321–325.
- Phatak, A. A., Wieland, F. G., Vempala, K., Volkmar, F., & Memmert, D. (2021). Artificial intelligence-based body sensor network framework—narrative review: Proposing an end-to-end framework using wearable sensors, real-time location systems and artificial intelligence/machine learning algorithms for data collection, data mining and knowledge discovery in sports and healthcare. *Sports Medicine-Open*, *7*(1), 1–15.
- Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *Springerplus*, *5*(1), 1–13.
- Tovar, J. (2021). The debate of VAR. In *On fairness, justice, and VAR* (pp. 29–39). Palgrave Macmillan.
- Wright, R. E. (1995). *Logistic regression*.
- Wunderlich, F., & Memmert, D. (2016). Analysis of the predictive qualities of betting odds and FIFA world ranking: Evidence from the 2006, 2010 and 2014 football world cups. *Journal of Sports Sciences*, *34*(24), 2176–2184.
- Wunderlich, F., & Memmert, D. (2018). The betting odds rating system: Using soccer forecasts to forecast soccer. *PLoS One*, *13*(6), e0198668.



Time Series Data Mining

Rumena Komitova and Daniel Memmert

Contents

- 17.1 Example Sport – 142**
- 17.2 Background – 143**
- 17.3 Applications – 144**
 - 17.3.1 Tasks in Time Series Data Mining – 144
 - 17.3.2 Time Series Data Mining in Medicine – 145
 - 17.3.3 Time Series Data Mining in Sports – 145
- References – 147**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Time series data are common in the sport domain and require often particular analysis techniques.
- Extracting information about the behavior of time series data is potentially very important and useful for the analysis of sport data.
- Time series data mining techniques are a good candidate for the application on sports data in order to discover unknown knowledge.
- Time series data mining techniques are useful for decision making in the sport domain by looking for the existence of patterns, discovering of motifs, or detecting anomalies.

17.1 Example Sport

In soccer, performance analytics based on positional data (the movement of the players and the ball on the pitch over time) and event data (certain meaningful information during a match) about the match are beneficial for the success of a team (see ► Chaps. 5 and 6). In many cases positional data can be considered as time series data. Furthermore, there are related efforts in linking the time series and sequence of events to each other. One way to convert a time series to an event sequence is by detecting multiple events such as player activities from the time series data. Activity recognition in soccer can be done thus by using the player and ball positional data. Broadly speaking, sport activity recognition in context of time series data the problem of discovering, and locating meaningful actions from time series representing different activities. For example, pass event can be defined as the moment in which the ball leaves the foot of the player. In tactical scenarios, the availability of time series data about player's activities along with their locations could be beneficial for their performance, which could be also helpful to support decision making in training. Furthermore, by the possibility to recognize all actions

of a soccer game only by means of position data in form of time series, it is possible to recognize the actions of players without the help of video-data annotations.

17.2 Background

Data mining is a field in computer science to discover and extract useful information from the data. Although, traditionally, analysis of sport data is based on expert knowledge and statistical analysis, data mining techniques have become increasingly popular in the field (Stein et al., 2017). The evolution of tracking systems becomes a further opportunity for researchers in the sport domain to extract new knowledge related to player performance, and movement patterns, among others. With the amount of information in the sport domain opportunities for data mining can be extremely widespread, and benefits from the results can be enormous (Bonidia et al., 2018; Ofoghi et al., 2013). For example, recurring and surprising events extracted from sport data could be helpful for sport activity recognition and patterns could be helpful in forecasting of future events, such as prediction of performance, coaching, and strategy planning.

There exists a large number of data mining applications and domains where information is reordered over a period of time, leading to sequence of temporal data (also called samples or observations). For example, in the sport domain, one major source of temporal data generated is coming from sensor data, which can be represented as *time series*. Application of time series techniques in data mining is called *time series data mining* (TSDM) (Esling & Agon, 2012; Fu, 2011; Mitsa, 2010; Komitova et al., 2023). Traditional time series analysis (Box et al., 2016) contains methods to analyze time series data in order to extract temporal rules from the structure of time series, such as trends, changes in value, seasonality, periodicity, or other characteristics of the data to generate an accurate forecast. TSDM, in contrary, deals with much larger amounts of time series data and much higher number of time series. However, the focus of TSDM is less on the analysis of the statistical properties of time series data, but instead focuses on the discovery of hidden relations between time series and extract potentially useful and meaningful information from them, where the terms “useful” and “meaningful” depend on the application. TSDM addresses tasks such as *classification* and *clustering* of time series, *anomaly detection*, and *motif discovery* in time series, and some more.

While the sport domain is only mentioned barely, the use of TSDM tasks could also be more extended to sport. State-of-the-art approaches used in medicine and individual sports are hardly applicable. For example, multiple types of human motion can occur within a recording session of physical activity (Minnen et al., 2006; Tanaka et al., 2005). TSDM tasks such as anomaly detection and motif discovery for human motion rely on similarity between (single or multidimensional) time series sequences of activities in form of motifs. Similar activities are characterized by similar sets of actions that appear frequently in sports, too.

The activities of a player can be broadly classified into simple activities (or events) and complex activities. Simple activities do not depend on the context, i.e.

they can exist by themselves (e.g. specific actions made by individual players). Complex activities on the other hand are composed of a set of simple activities and may focus on understanding the relationship between the event, or the interactions between other players and analyzing them as a sequence of events. For example, time series data from the position data of a player during a match can be broken down to events as passes, or shots but even more complex activities such as dribbling can be regarded as combination of multiple single activities. Activity recognition (or event detection) in sports by considering positional data in form of time series can thus provide valuable knowledge and context about the actions of a player. However, such a task is not a simple task because there is no standard taxonomy of player activities. Additionally, it is not easy to model complex activities, or context category precisely and generally enough.

Definition

A *time series* T is an ordered sequence of real numbers, i.e. $T = [t_1, \dots, t_n]$, where $t_i \in \mathbb{R}$, $i = 1, \dots, n$, denotes the i -th element of the time series T . A time series can be a collection of observations from one source, i.e. one sensor. A *multidimensional time series* is a set of single time series. With regards to sensor data, the term multidimensional implies the exploration of multiple time series (signals) in parallel. *Time series data mining (TSDM)* is a field in data science to discover and extract useful information from time series data.

17.3 Applications

17.3.1 Tasks in Time Series Data Mining

► Example 1

Time series *classification* and time series *clustering* is an important and challenging problem in time series data mining (Esling & Agon, 2012; Liao, 2005; Mitsa, 2010). Time series classification seeks to assign labels to each time series of a set. Given an unlabeled time series, the goal of time series classification is to assign it to one out of a given number of predefined classes. Clustering is method of creating natural groups, so called *clusters*, in a dataset. Input of the clustering can be a set of time series (multidimensional time series) from different sources, such as sensors, or the set of subsequences of a single large time series data source. Considering a set of time series, the main idea of clustering is to find groups of time series that are similar inside the cluster but are relatively different from time series of other clusters. ◀

► Example 2

Time series *anomalies* can be defined as unexpected or unusual patterns in time series data that do not conform to a well-defined notion of the expected (normal) behavior (Zolhavarieh et al., 2014). In other words, anomalies appear when the underlying process deviates from its normal behavior. The problem of finding them is referred to as

anomaly detection. Anomalies can be broadly classified into three general categories (Chandola et al., 2009). A *point anomaly* is a point that deviates significantly from all the points in the dataset (Braei & Wagner, 2020). *Contextual anomalies* are data points whose values are anomalous with respect to a specific context, but not otherwise. That is, a given behavior might be “normal” in concrete context but abnormal on another. Finally, a *collective anomaly* refers to a collection of related data instances that individually may not be anomalies, but their collective appearance is anomalous. Knowing a priori which type of anomaly the time series data might contain, helps the data analyst to choose the appropriate detection method. ◀

▶ Example 3

A common problem in the time series data mining and machine learning community is the finding of previously unknown, frequently occurring subsequences of single (or multiple) time series, also called *motifs* (Chiu et al., 2003; Lin et al., 2002). *Motif discovery* is the technique to find them and is a fundamental problem for time series data mining (Mueen, 2014; Torkamani & Lohweg, 2017; Tanaka et al., 2005). Examples for motifs can be peaks (e.g. local minima or maxima), changes of noise characteristics of time series, or a variation of time or spectral components, which repeatedly occur in a time series. However, finding motifs is a difficult task, even when they have the same or very similar general characteristics, because in most cases the number of occurrences of motifs, their shape and length, and duration of occurrences may be unknown (Mitsa, 2010). ◀

17.3.2 Time Series Data Mining in Medicine

▶ Example 4

Anomaly detection and motif discovery in the medical domain is a very critical problem with requires high degree of accuracy (Lin & Li, 2010; Liu et al., 2015; Sivaraks & Ratanamahatana, 2015). Electrocardiogram (ECG) is nothing but a time series which consists of the electrical impulses from the heart. Anomaly detection in ECG can help to detect the abnormal heartbeats before the diagnosis and motif discovery can help to locate the highly similar and rapid beats in ECG. For example, Wankhedkar and Jain (2021) proposed method to detect anomaly present in an ECG data. The beats similar to each other are the motifs whereas the beats having high value and dissimilar are the anomalies. ◀

17.3.3 Time Series Data Mining in Sports

▶ Example 5

Compared to the detection of heartbeats, which are often periodic, motifs (or events) such as passing, shooting, or dribbling in soccer do not regularly appear during soccer matches. Schuldhuis et al. (2015) provided a low-cost inertial sensor-based shot/pass approach for soccer teams to identify passes and shots on target, mainly consisting motif

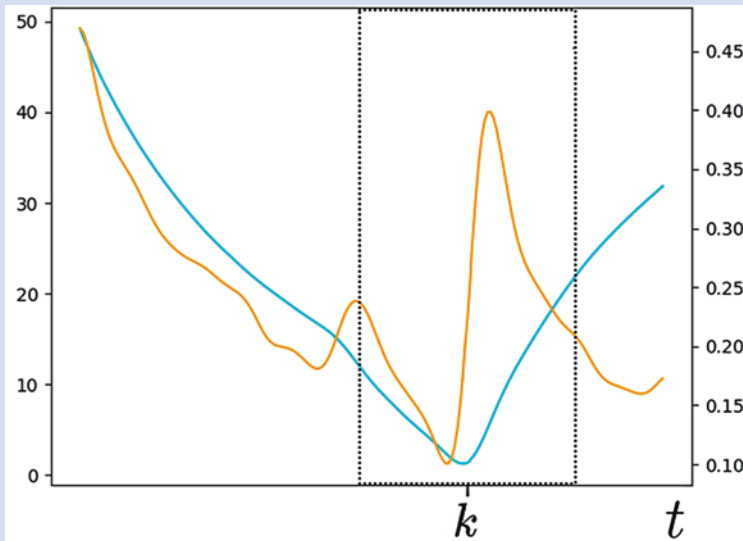
discovery and classical machine learning algorithms like support vector machine, classification, and Naïve Bayes. Peaks in the accelerometer data (motifs) has been detected from the left and right shoe of soccer player and the peaks has been classified regarding shot, pass, and other. The detection of *atomic* pass events, defined as the moment in which the ball leaves the foot of the player, was also conducted (Sanford et al., 2020). The authors propose machine learning algorithms that are capable of detecting passes, along with other group activities, from either the video or the position data. ◀

► Example 6

Yeh et al. (2017) for example used an algorithm to discover the correct (multidimensional) motif location of a boxer's punch (as repeated behavior). The algorithm matches a simple cross with the cross on a one to two combo, and the three dimensions: right upper arm, right forearm, and left upper lag. Two behaviors on the boxer's dominant hand are almost identical but is in a different position within different occurrences of the motif. ◀

Study Box

Biermann et al. (2023) used an algorithm to automatically detect events such as passes in the positional data of soccer matches using time series data mining techniques. To detect pass events the authors provided motif detection method from the player-ball distance and ball acceleration in form of time series obtained from the position data. Motifs occur when the time series data show a specific form or change at the same time (see ■ Fig. 17.1). The authors therefore suggested a method that used both time



■ Fig. 17.1 An exemplary sequence for player-ball distance (cyan curve) and ball acceleration (orange curve). Here, k denotes the time point at which a pass (an event) occurs

series from the position data and event data. Initially, they compared passes from the event data with the position data. Therefore, they instruct an expert to label passes in the position data. They found out that passes in the event data tend to be slightly delayed in comparison to the position data. To compensate this delay, they initially perform a time series motif detection method to construct a model that detects passes in the position data. To perform the detection of events in the time series data, the authors used an appropriate feature space representation of the time series. Subsequently, they take the (delayed) passes from the event data as a template and refine them in a certain window given the pass detection model. They report that this algorithm is largely improving the synchronization between position and event data.

? Questions for the Students

1. What is time series data mining (TSDM)?
2. Where can TSDM methods be used in sport science?

References

- Biermann, H., Komitova, R., Raabe, D., Müller-Budack, E., Ewerth, R., & Memmert, D. (2023). Synchronization of passes in event and spatiotemporal soccer data. *Scientific Reports*, *13*, 15878.
- Bonidia, R., Rodrigues, L., Avila-Santos, A. P., Sanches, D., & Brancher, J. (2018). Computational intelligence in sports: A systematic literature review. *Advances Human-Computer Interaction*, *2018*, 1–13.
- Box, G., Jenkins, G., & Reinsel, G. (2016). *Time series analysis: Forecasting and control* (5th ed.). Wiley.
- Braei, M., & Wagner, S. (2020). Anomaly detection in univariate time-series: A survey on the state-of-the-art.
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–58.
- Chiu, B., Keogh, E., & Lonardi, S. (2003). Probabilistic discovery of time series motifs. In *Proceedings of the 9th international conference on knowledge discovery and data mining (KDD)* (pp. 493–498).
- Esling, P., & Agon, C. (2012). Time series data mining. *ACM Computing Surveys (CSUR)*, *45*(1), 1–34.
- Fu, T.-C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*(1), 164–181.
- Komitova, R., Raabe, D., Rein, R., & Memmert, D. (2023). Time series data mining for sport data: A review. *International Journal of Computer Science in Sport*, *21*(2), 17–31.
- Liao, T. (2005). Clustering of time series data—A survey. *Pattern Recognition*, *38*(11), 1857–1874.
- Lin, J., Keogh, E., Lonardi, E., & Patel, S. (2002). Finding motifs in time series. In *Proceedings of the eighth ACM SIGKDD International conference on knowledge discovery and data mining 2nd workshop on temporal data mining* (pp. 53–68).
- Lin, J., & Li, Y. (2010). Finding approximate frequent patterns in streaming medical data. In *IEEE 23rd international symposium on computer-based medical systems (CBMS)*, IEEE (pp. 13–18).
- Liu, B., Li, J., Chen, C., Tan, W., Chen, Q., & Zhou, M. (2015). Efficient motif discovery for large-scale time series in healthcare. *IEEE Transactions on Industrial Informatics*, *11*(3), 583–590.
- Minnen, D., Starner, T., Essa, I., & Isbell, C. (2006). Discovering characteristic actions from on-body sensor data. In *Wearable computers, 2006 10th IEEE international symposium on wearable computers*, IEEE (pp. 11–18).
- Mitsa, T. (2010). *Temporal data mining*. Chapman and Hall/CRC.

- Mueen, A. (2014). Time series motif discovery: Dimensions and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(2), 152–159.
- Ofoghi, B., Zeleznikow, J., MacMahon, C., & Raab, M. (2013). Data mining in elite sports: A review and a framework. *Measurement in Physical Education and Exercise Science*, 17(3), 171–186.
- Sanford, R., Gorji, S., Hafemann, L. G., Pourbabae, B., & Javan, M. (2020). Group activity detection from trajectory and video data in soccer. *Proceedings of the IEEE/CVF conference on computer vision, graphics and image processing* (pp. 1–7).
- Schuldhuis, D., Zwick, C., Körger, H., Dorschky, E., Kirk, R., & Eskofier, B.M. (2015). Inertial sensor-based approach for shot/pass classification during a soccer match. In *KDD workshop on large-scale sports analytics* (pp. 1–4).
- Sivaraks, H., & Ratanamahatana, C. (2015). Robust and accurate anomaly detection in ECG artifacts using time series motif discovery. *Computational and Mathematical Methods in Medicine*, 2015, 1–20.
- Stein, M., Jenezko, D., Seebacher, D., Jäger, A., Negel, J., Hölsch, M., Kosub, S., Schreck, T., Kleim, D., & Grossniklaus, M. (2017). How to make sense of team sport data: From acquisition to data modeling and research aspects. *Data*, 2(1), 2.
- Tanaka, Y., Iwamoto, K., & Uehara, K. (2005). Discovery of time series motif from multidimensional data based on MDL principle. *Machine Learning*, 58(2), 269–300.
- Torkamani, S., & Lohweg, V. (2017). Survey on time series motif discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2), e1199.
- Wankhedkar, R., & Jain, S. K. (2021). *Motif discovery and anomaly detection in an ECG using matrix profile*. *Progress in advanced computing and intelligent engineering* (pp. 88–95). Springer.
- Yeh, C., Kavantzias, N., & Keogh, E. (2017). Matrix profile VI: Meaningful multidimensional motif discovery. In *IEEE international conference on data mining (ICDM)*, IEEE (pp. 565–574).
- Zolhavarieh, S., Aghabozorgi, S., & Teh, Y. (2014). A review of subsequence time series clustering. *The Scientific World Journal*, 2014, 312521.



Process Mining

Marc Garnica Caparrós

Contents

- 18.1 Example Sport – 150**
- 18.2 Background – 151**
- 18.3 Application – 153**
 - 18.3.1 Process Mining in Healthcare – 153
 - 18.3.2 Process Mining in Education – 153
 - 18.3.3 Process Mining in Soccer – 153
- References – 154**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Process Mining is a discipline that derives from Data Mining and focuses on the analysis of end-to-end processes usually represented by execution logs.
- Execution logs are a time-ordered collection of events occurring in any system. For instance, the procedures a patient needs to overtake in a hospital.
- Process Discovery is the set of algorithms that conceptually define and visualize a process from collected data (observed behaviour).
- Conformance checking measures the alignment between the process design (theoretical behaviour) and the actual execution observed in the data.
- Process enhancement studies the process and its conceptual definition with the aim to improve the overall system.
- Process Mining is highly applicable to any field and opens several opportunities for sports science, for instance, using event and positional data from invasion sports.

18.1 Example Sport

Basketball is a team sport although the media tries to individualise certain actions or players. Basketball teams train weekly to practice their offensive and defensive tactics. Often, these tactics might vary depending on the opponent. Analytics based on positional data (the trajectories of the players and the ball on the court) and event data (the actions performed by the players) gives tactics coaches an overview of how a team is behaving on the court. In attack, teams usually have predefined offensive plays, thus, players know how to start the attack, move the ball and look for scoring opportunities. These predefined attack systems can be considered a theoretical process with a start, a set of activities and resources and several end options. These processes are also present in the collected data, on-ball actions are present in the event data while off-ball actions are included in the positional data. Process Mining provides the toolset to discover these processes in the collected event and positional data, compare the observed team behaviour to the theoretical

design done by the coaches and finally try to improve these processes, for instance, by identifying ignored scoring opportunities in the collected data. The implementation of Process Mining could be beneficial for sports performance and support decision-making in sports tactics.

18.2 Background

The field of Process Mining (PM) is derived from data mining, a research discipline involving mathematics, statistics and computer science for knowledge extraction from datasets. Often, these datasets deal with big data challenges such as large volume, variety in their syntax and semantics or high speed of generation. In the 1990s, the implementation of numerous IT systems was being conducted in all areas of the modern industry revolutionizing every business. The evolution of business digitalization brought together software systems such as Enterprise Resource Planning (ERP) and Business Process Management (BPM) tools and methodologies. In these early days, business process definition was part of the design phase of any enterprise, experts would identify the requirements and design the most efficient process to conduct a certain task involving numerous resources, activities and agents (Tiwari et al., 2008).

Along with data mining, several other approaches started surfacing as a way to deal with such complex implementations. Data warehousing, database management system and advances in hardware allowed the massive collection of all the core aspects of any business process, annotating every step and every resource used during the process in the so-called execution log. Motivated by this massive collection of business processes, the work of (van der Aalst et al., 2004), presented the idea of creating data-driven *process models* or *workflows* that can be observed, analysed and improved directly from the collected data. Therefore, Process Mining makes use of the execution logs of any system to reconstruct the actual business processes.

Although these execution logs, often called event logs, are omnipresent in so many industries, organizations lack a good definition and understanding of their actual processes. Thus, PM untangles the differences between event logs (observed behaviour) and process models (either theoretical models on how systems should work, or data-driven discovered models of system behaviour). The applications of such analysis include but are not limited to analysing the treatment of patients in hospitals, improving user experience on e-commerce websites, analysing baggage management in an airport system or controlling an automatic industry of a car manufacturer. In fact, while the main work on PM is still part of the academic knowledge, some businesses started applying this analytical technique to fully understand and control their processes.

It is not until 2011 that the Process Mining Manifesto (van der Aalst et al., 2012) is presented to the research community. In this document, PM is defined as a technique to extract knowledge from event logs generated by information systems with the end goal to discover, monitor and improve processes in various business domains. In this work, it is also stated the three main PM types: discovery, conformance and enhancement. *Process discovery* is the most research prolific type, discovery techniques cover

the generation of processes and visual representations directly from event logs without additional information (van der Aalst, 2016), it allows for an automated definition of your business process from your collected data. *Conformance checking* techniques analyse the deviations between the data-driven automatically discovered models with the theoretic process (how I process should work) (Rozinat & van der Aalst, 2008; Bergami et al., 2021). Finally, the *process enhancement* methodologies aim to improve the existing processes using the information from the event logs by modifying the activities or reorganizing the resources involved (de Leoni, 2022).

The existence of such event logs in a system not always needs to be a consequence of an IT infrastructure in place. In some cases, event logs are generated from sensor data observing a natural system (e.g., weather forecast stations) and the process is actually a conceptual definition trying to describe the observed behaviour. In these cases, the end goal of PM can still be applied. In sports, several data collection techniques are being executed on all kinds of levels. Professional sports teams and federations monitor their athletes' heart rate, distance covered or average fatigue levels thanks to sensor devices. In sports included under the umbrella of invasion sports (Hughes & Bartlett, 2002), games are also monitored generating two main data sources: event data and positional data. Event data is originally used to generate game statistics, counting the number of actions of a certain player, reporting the accuracy of passes etc. However, event data is by nature the execution log of a team in the game, the timely ordered sequence of all actions occurring in the game. Similarly, positional data collects the position of all players and the ball at a high frequency throughout the game. While this data is by definition a spatiotemporal series of positions, discrete methods could produce a sequence of positional phases, trajectories or movement patterns that could fit under the definition of a process. In both cases, PM could consume these data sources to methodologically define the process underneath a team's performance in a game. Research in this field could elaborate on how to use PM techniques to define team tactics or player functions.

Definition

Process mining (PM) is a relatively young research discipline in the intersection of data mining and business management that conducts discovery, conformance and enhancement of business processes. Process-aware analysis manages operational studies assuming that the data is based on a dynamic behaviour process. Roles and actors interact towards a certain goal or function in time and execute certain patterns, orders, and workflows. Some data mining techniques, such as sequence or episode mining, can model data as a sequence but not consider end-to-end processes. PM identify the process models and provides a visual representation of structure or unstructured event logs (Diamantini et al., 2016). Process discovery involves the identification of process models and visual representations of structure or unstructured process data sourcing from any environment. Conformance checking provides a measure of alignment between the strategy of a process (the theoretical behaviour) and the actual execution. Finally, process enhancement extends or improves the actual process models using knowledge extracted from the analysis.

18.3 Application

PM is a highly applicable area and has proven to be extremely insightful in all kinds of domains and has become a core aspect of any system implementation.

18.3.1 Process Mining in Healthcare

► Example 1

Hospital emergency rooms are a complex system where fast decision-making is required. Indicators such as waiting times, patient congestion or quality of care are always a priority. PM has proven to be useful to observe and analyse the interactions, activities and processes that are undertaken in an emergency room (Rojas et al., 2019). The purpose of this recent study was to use PM as a tool for performance analysis in this area as well as to identify bottlenecks in the whole procedure that eventually could be avoided to improve the quality of the healthcare service. ◀

► Example 2

Another great application of PM in healthcare was showcased in a hospital behaviour analysis research work (Arnolds & Gartner, 2017). In this study, the clinical pathways, the set of procedures that a patient undergoes in the hospital, were the centre of the process-aware analysis. The prediction of certain clinical pathways such as diagnostics, surgery or therapy and the transitions between them is of great importance for hospital layout planning. Through the analysis, the authors achieved a reduction in the distances travelled by the patients. ◀

18.3.2 Process Mining in Education

► Example 3

Since 2020, online communication and collaboration have become an essential part of any team or organization. Indeed, even educational systems are moving towards a more flexible service model combining online resources with face-to-face classes. In a study from 2016 (Alvarez et al., 2016), PM was used to measure the alignment between teacher goals and student activities in an e-learning environment. The authors made use of the logs generated by the students when doing their activities and the teacher's course plan (observed vs. theoretical behaviour). ◀

18.3.3 Process Mining in Soccer

► Example 4

PM has actually already been tested in sports, in this case using soccer event data (Kröckel & Bodendorf, 2020). In this exploratory study, the opportunities of PM were analysed as a tool for the tactical analysis of soccer games. PM shows promising potential to evalu-

ate team and player's performance in soccer and provide a standard definition of team tactics by detecting the typical behaviour collected in the event data. PM also allows investigating the organizational perspective of a team's style of playing. Thus, extracting metrics on how players are interacting with each other towards a common goal. ◀

Study Box

Invasion sports are sports disciplines sharing core concepts on their way of playing and how the confrontation between the two teams is executed. In sports like soccer, basketball, handball or rugby; teams share the purpose to invade the opponent's territory and scoring points while keeping the opponent's point as low as possible in a defined time period. Despite the similarities, research on invasion sports is mainly presented in one-dimensional studies analysis of a single sport. The reason for this discretization of the studies might be related to data availability or a more easy formulation of the study requirements. In this scenario, Process Mining raises as a candidate to offer a sport-independent analysis where process models could be generalized to any sport. For instance, abstracting the common characteristics between a soccer goal and a rugby try. Understanding general concepts of invasion sports in terms of location-based control of the field or the collaboration between players towards common tasks could open up new tactics in the respective sports and substantially contribute to sports science. In fact, recent studies have noted the positive effect of multi-sport practices on young athletes for their potential long-term development and specialization (Barth & Güllich, 2020).

? Questions for the Students

1. Where do we have examples of execution logs in sports analysis?
2. Define the three types of Process Mining.

References

- Alvarez, P., Fabra, J., Hernandez, S., & Ezpeleta, J. (2016, September). Alignment of teacher's plan and students' use of LMS resources. Analysis of Moodle logs. In *2016 15th international conference on information technology based higher education and training (ITHET)*. IEEE. <https://doi.org/10.1109/ithet.2016.7760720>.
- Arnolds, I. V., & Gartner, D. (2017). Improving hospital layout planning through clinical pathway mining. *Annals of Operations Research*, 263, 453–477. <https://doi.org/10.1007/s10479-017-2485-4>
- Barth, M., & Güllich, A. (2020). Non-linear association of efficiency of practice of adult elite athletes with their youth multi-sport practice. *Journal of Sports Sciences*, 39, 915–925. <https://doi.org/10.1080/02640414.2020.1851900>
- Bergami, G., Maggi, F. M., Marrella, A., & Montali, M. (2021). Aligning data-aware declarative process models and event logs. In *Lecture notes in computer science* (pp. 235–251). Springer International Publishing. https://doi.org/10.1007/978-3-030-85469-0_16
- de Leoni, M. (2022). Foundations of process enhancement. In *Lecture notes in business information processing* (pp. 243–273). Springer International Publishing. https://doi.org/10.1007/978-3-031-08848-3_8

- Diamantini, C., Genga, L., & Potena, D. (2016). Behavioral process mining for unstructured processes. *Journal of Intelligent Information Systems*, 47, 5–32. <https://doi.org/10.1007/s10844-016-0394-7>
- Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of Sports Sciences*, 20, 739–754. <https://doi.org/10.1080/026404102320675602>
- Kröckel, P., & Bodendorf, F. (2020). Process mining of football event data: A novel approach for tactical insights into the game. *Frontiers in Artificial Intelligence*, 3, 47. <https://doi.org/10.3389/frai.2020.00047>
- Rojas, E., Cifuentes, A., Burattin, A., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2019). Performance analysis of emergency room episodes through process mining. *International Journal of Environmental Research and Public Health*, 16, 1274. <https://doi.org/10.3390/ijerph16071274>
- Rozinat, A., & van der Aalst, W. M. (2008). Conformance checking of processes based on monitoring real behavior. *Information Systems*, 33, 64–95. <https://doi.org/10.1016/j.is.2007.07.001>
- Tiwari, A., Turner, C. J., & Majeed, B. (2008). A review of business process mining: State-of-the-art and future trends. *Business Process Management Journal*, 14, 5–22. <https://doi.org/10.1108/14637150810849373>
- van der Aalst, W. (2016). *Process mining*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-49851-4>
- van der Aalst, W., Weijters, T., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, 16, 1128–1142. <https://doi.org/10.1109/tkde.2004.47>
- van der Aalst, W. M. P., Adriansyah, A., Alves De Medeiros, A. K., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., van den Brand, P. C. W., Brandtjen, R., Buijs, J. C. A. M., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., de Leoni, M., et al. (2012). Process mining manifesto. In *Business process management workshops* (pp. 169–194). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28108-2_19



Networks Centrality

*João Paulo Ramos, Rui Jorge Lopes, Duarte Araújo,
and Pedro Passos*

Contents

- 19.1 A Network Science in Football – 158**
- 19.2 Background – 159**
- 19.3 Applications – 162**
- References – 166**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Performance-relevant parameters from training and competition can be captured by network science
- Network Analysis (NA) help to understand performance outcomes
- There is a distinction between dynamics of the networks vs. dynamics on the networks (structure vs. processes).
- There is a distinction between match-level and play-level metrics.

19.1 A Network Science in Football


The dynamics of players' interactive behavior implies that the relevant moments of a match evolve over time. The nature of the interactions among team players' and opponents are a complex process, that goes beyond passes, positioning and distances between players. To capture such complexity, hypernetworks approach is a promising tool. Previous work on performance analysis (PA) in team ball sports has focused on centrality. In the last two decades, dynamical systems theory (Davids et al., 2003, 2005; Araujo et al., 2004; Reilly et al., 2005) and later, ecological dynamics have been used as a background to describe and explain interactive behaviors in team sports (Vilar et al., 2012). Consequently, questions about cooperative interactions between players from the same team emerged, for instance, who is the player that interacts the most with their teammates? Additionally, the question of centrality within cooperative and competitive interactions increases relevance leading to more specific research questions such as: Considering players' relative positioning and distance, what are the sets of players (1 vs. 1; 2 vs. 1; 1 vs. 2; 2 vs. 2; etc. ...) that are more frequent (central) in the matches? Where do they occur? Who are the players involved? or what are the processes used by the players and teams to promote local dominance (e.g. 2 vs. 1) in the different areas of the pitch or gain advantage in goal scoring opportunities?

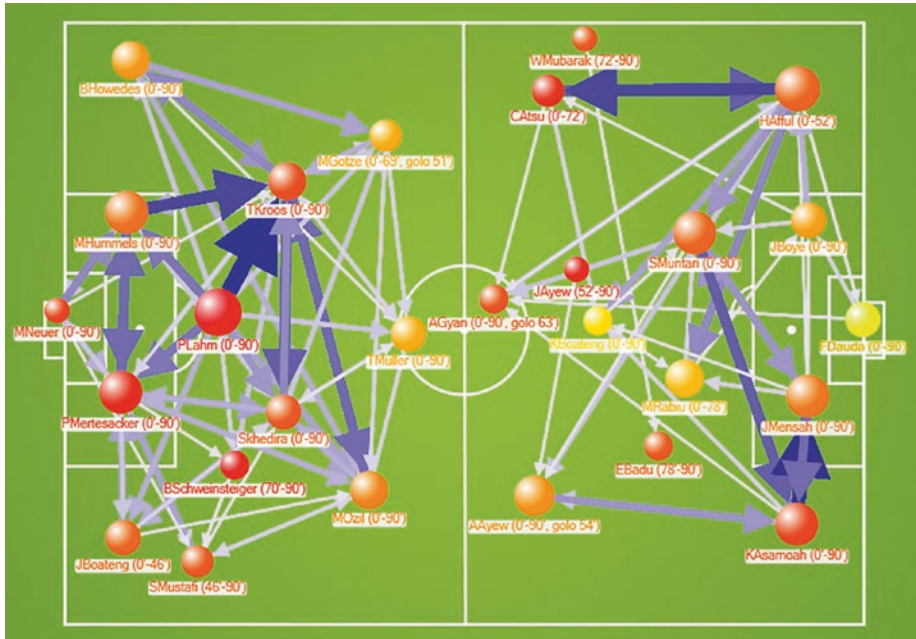
19.2 Background

Complex systems approach to team sports performance via network science considers and describes teams as having their parts (groups and players) interconnected (Bar-Yam, 1997). Therefore, the level of system complexity can be made higher, through structure by increasing the number of their parts and, via functionality by increasing their interactions (Bar-Yam, 2004). In team sports, the number of players (parts) in a match is regulated, which highlights the complexity of the amount and variability of interactions between groups and players (Bar-Yam, 2003).

With the aim of studying interactions, an increasing number of studies (Duch et al., 2010; Dutt-Mazumder et al., 2011; Passos et al., 2011; Duarte et al., 2012; Grund, 2012; Clemente et al., 2014, 2015a, 2015b, 2015c; Ramos et al., 2017a, 2017b) used Social Network Analysis (SNA) as a suitable method to address the interdependencies in team sports. The approach consisted in modeling the cooperative interaction between dyads of players from the same team or their actions with the ball. The communicative power of SNA visualization by representing the aggregated passing data across a football match (Ramos et al., 2018) between the dyads, attracted the use of these metrics to performance analysis towards modeling intra-team coordination as the frequent passing interaction between players in team sports (Duch et al., 2010; Passos et al., 2011; Brandes et al., 2012; Grund, 2012).

The analysis of these passing networks reveals some collective properties (e.g., patterns) of team performance as well as the underlying individual contribution to those properties (Ramos, 2019). However, for practitioners it is still unclear in what way social networks analysis contributes to an effective individual and team performance improvement. Some uncertainty remains concerning how individual differences and roles (influential players) are combined to enhance team performance outputs (Duch et al., 2010). A chief principle present in every network analysis of a system or process assumes that there is a structure, i.e., a network, in which these interactions occur (Brandes et al., 2013).

We now turn to the structural properties of networks, i.e., to the properties that are common to all networks. Given a particular system of study, there are properties that are expressed via network science. First and foremost, the system is composed of interacting elements, i.e., a network is a “collection of vertices joined by edges” (Newman, 2010) as represented in  Fig. 19.1. In the many different approaches to study team ball sports, players and passes are represented respectively by the vertices (actors) and edges (interactions) of the network. In the simplest network models (simple networks or simple graphs) there is at most one edge between any pair of vertices, whereas more sophisticated models (multigraphs) may have multiple edges between pairs of nodes (Ramos, 2019). The most basic global structural properties of a network are order and size, expressed in its number of vertices and edges. The global intensity level of interactions on the network can be gauged via the density property computed as the relation between the network size (number of passes or other actions with the ball—edges) over the maximum number of edges that could possibly exist within the network (Guillaume &



■ **Fig. 19.1** Germany—Ghana: FIFA World Cup 2014. Each circle represents a player in his relative position; the radius and colour of each circle represents the number of players that a player interacts with and his pass precision (red more precision; yellow less precision), respectively; the arrows represent the direction of the passes between players; the width and shade of each arrow represents the number of interactions (passes) between players (lighter arrows indicate less passes, darker arrows indicate more passes). The numbers in brackets represent the minutes played by each player (e.g., $-90'$, played 90 min) or the moment in the match when a player started to play (e.g., $+78'$, entered in the match at minute 78 and *golo 54'* means that scored a goal at minute 54'

Latapy, 2006). High density in the network tends to be associated to better team performance (Grund, 2012; Ramos, 2019).

The referred metrics assesses the dynamics on the network, focusing on flows across the network structure, e.g., ball passes between players. In this way, an entire match is represented by the aggregate of all the passes that occurred, as static flip books (cumulative snapshots of the network as a function of time) of each player action, where player's position remains constant but interactions cumulate over time (Moody et al., 2005; Ramos et al., 2018). The structural properties studied in the literature were mainly, network centrality and the density of interactions between team members and the performance outcomes (Katz et al., 2004; Balkundi & Kilduff, 2006; Grund, 2012). By counting the number of edges (e.g., passes) connected to a vertex (e.g., player), the degree of a vertex is obtained, which is also called degree centrality (Newman, 2010). The studies on team ball sports have been mostly looking at the “ball flux” (e.g., ball passes) which form directed networks from one player to another. On this sort of networks the vertices (players) have two degrees: the in-degree (e.g. the number of ingoing passes to that player, or intercept-

tions made) and the out-degree (e.g. the number of the outgoing passes or interactions conceded) (Newman, 2010). If one team has few players with high centrality values, it means that those players are responsible for most of the passes performed and team tends to be highly dependent on them, with a more predictable behavior, which is associated to a less efficient team performance (Grund, 2012; Ramos, 2019).

By counting the number of edges and the number of vertices, the intensity of the network is computed. Moreover, given the relation between the network size (number of passes or other actions with the ball—edges) over the number of edges that could possibly exist within the network, the density of the network is calculated (Guillaume & Latapy, 2006). High density in network tends to be associated to a better team performance (Grund, 2012; Ramos, 2019; Pina et al., 2017). The referred metrics assesses the dynamics on the network, focusing on flows across the network structure, e.g. ball passes between players. Centrality studies could help practitioners improving team sports performance by answering questions like: “who is the most interactive player?” or “which players have an intermediary role?” or “how central is a player?” or “how does each player contribute to the performance of the others?” (Ramos et al., 2018).

Tackling the question about the most interactive player, the focus is on the local structural analysis of the interactions. That is, on the interactions/passes between the player and the adjacent players are given by the in-degree or the out-degree. To tackle other questions, we have to understand the global analysis of network structure because it considers not only the adjacent vertices but that a second, a third and/or more steps occurred in those sequences of passes. One limitation of using these metrics which are imported from other complex systems networks analysis, like Internet (social) networks, is that some metrics are based on the concept of shortest path or geodesic path. This concept refers to the path between two vertices for which there is no other path in the network that is shorter (is the geodesic distance or shortest distance that represents the shortest network distance) between those two vertices (Newman, 2010). When analyzing soccer matches, we have the aggregate of all passes/actions for the entire match or the aggregation during relevant time spans or between some significant events (e.g., goals or goal scoring opportunities). In these cases, we need to consider a more global network analysis, which includes all the connections that occurred during the course of a match or time span. That aggregating feature in team ball sports, does not follow the shortest path concept (typical in Internet social interactions), but instead the concept of random walks (a walk that takes random steps across the network, e.g. one player could be a part of the sequence of passes, for more than once). This concept represents better the events during a team sport match (Ramos et al., 2017a, 2017b), such as ball passing (Newman, 2010). Metrics like betweenness centrality, that expresses the degree in which one vertex lies on the shortest path between two other vertices, or closeness centrality, which is a measure of centrality that considers the length of the shortest paths between the focal vertex and all the other vertices. This metric is able to help answering questions about the intermediate and central players, respectively, but it assumes that interactions must happen through the shortest paths (Ramos, 2019).

Definition

Network Analysis (NA) starts with a theory about the system (Brandes), the existence of an underlying network in the system is a chief assumption of this theory (e.g., a social theory in the case of Social Network Analysis). NA is the computation of network or node properties (such as network order or node degrees) which is only a part of the broader process of network science.

Flow centrality (Freeman, 2000) can overcome the limitation of *betweenness centrality* (basis on shortest paths), because it counts the fraction of walks that leads to an event where the focal player is involved, rather than considering the shortest paths between players (Duch et al., 2010). Fewell et al. (2012) applied it in basketball research where it can assess the individual dominance on play-level by focusing on the overall involvement during all plays in a match (Fewell et al., 2012). By calculating *flow centrality*, it is possible to capture the involvement of each play position in all plays across a match. Building on this metric and *random-walk betweenness* (Newman, 2005), it was also possible to compute a new metric called *flow betweenness*, which measures the fraction of plays in which the focus player functions as an intermediary player relative to all plays by its team (Korte et al., 2019). Moreover, regarding betweenness, the *weighted betweenness* scores can be calculated for each playing position and assesses how often a player is in-between any other two players of its team measured by their strongest passing connections (or other variables of interaction) across a match, thus the player functions as a bridging unit within plays. These metrics can be used to identify the playmakers and answer the question about the intermediary player, when substituting match-level (aggregate of all the plays in the match) with play-level metrics (the aggregate of some instants before a specific event in the match, e.g. a goal scoring opportunity). Additionally, a distinction between play-level metrics is necessary. These metrics emphasize different tasks among playing positions. Ramos et al. (2018) first suggested that *flow centrality* might be a suitable playmaker indicator that highlights intermediary players on play-level (Ramos et al., 2018).

19.3 Applications

► Example 1

Whereas most of the studies with SNA in sports focused on the attacking patterns, Sasaki and colleagues (Sasaki et al., 2017) aimed to clarify the networks created within the defensive patterns that play a decisive role during a Rugby match. The rationale was that the cooperation that links structural entities (e.g., defenders) is always dynamic. Thus, a different collaborative format indicates different characteristics of an entity. Again, it was expected that the more interactive structure (higher complexity) is less predictable and more adaptable. ◀

► Example 2

A few studies have investigated cooperation through direct physical contact, e.g., tackling in rugby (Koh et al., 2013). The authors calculated the absolute frequency of tackling per position during a rugby match, which contributes to a defense turnover performance, which consequently neutralized the offensive activity. The purpose of this study was to explain the centrality of a defensive squad in rugby. Whenever a multi-player defense act occurred, a singular network was created. The vertices represented the positions of the players, and the edges represented cooperation between teammates in the course of a match (double tackling), when a tackle with two players led to a turnover in play. The edges were weighted by the number of repetitive cooperative actions in the aggregation of the entire match. ◀

► Example 3

There still exists a gap between SNA and performance outcomes that fosters the practical impact of the approach. The current temporal approaches did not consider the actual sequence of ball passing to detect players that are in fact connecting their team members through passing. This implies that passing sequences should be evaluated separately instead of examining the aggregated passing data across a match (Ramos et al., 2018). Thus, the interplay in each ball possession needs to be analyzed separately instead of evaluating an aggregated passing matrix at match level. This type of approach tackles the *dynamics of the network* assessing the changes in the network structure itself, like the: (1) relational space (i.e. interactions considered in a geographical space); (2) their time structure (i.e., rate of change, order or sequence, or simultaneity of interactions); and (3) their relations with different types of vertices (i.e. teammates or opponents), thus considering both cooperative and competitive interactions (Moody et al., 2005; Ramos et al., 2018), including the opportunity for a pass between ball carrier and her/his team mates was assessed, via the possibility that such pass may or not be intercepted by an opponent player. This landscape of passing affordances can be represented as edges between players and/or spatial locations on the pitch (Passos et al., 2020). ◀

► Example 4

Complex networks like temporal and bipartite, such as the so-called hypernetworks (Johnson, 2006) represent interactions and relations that occur during the course of a team sports match. In a hypernetwork a hyperedge can connect more than two nodes (e.g. two players from one team and one from the other team: 2 vs. 1), directly representing n -ary relations as sets, σ (Johnson, 2006, 2008, 2013, 2016; Criado et al., 2010; Boccaletti et al., 2014; Ramos et al., 2017a, 2017b). ◀

1. This generalization enables the representation of multiple cooperative and competitive interactions in their exact positioning and time frame on the match (■ Fig. 19.2). The resulting representations and statistics can describe:

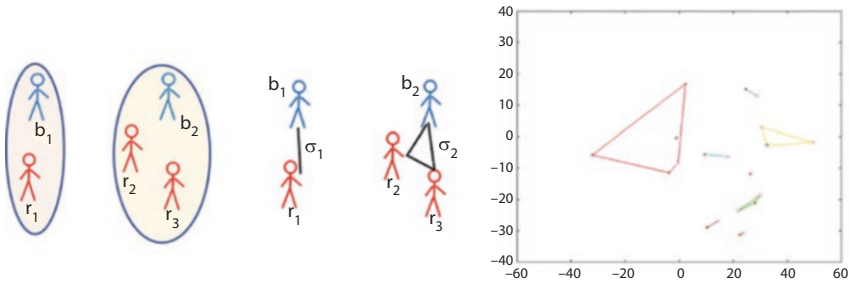


Fig. 19.2 Hypernetwork construction based on proximity between players' forming sets of players connected through edges (polygons) and represented in their position in the soccer pitch on a given instant of time

2. If there are central sets of each type (e.g. 1 vs. 1; 2 vs. 1; 1 vs. 2; 2 vs. 2) that occurs in the match or in some time spans (e.g. the 10 s before a goal scoring opportunity);
3. Who are the central players in the sets and if there are a specific central set that occurs more than the others;
4. If there are a central specific region where the sets occur, represented in histograms like *heat maps*;
5. How sets disaggregate or aggregate (e.g. velocity changes in players moves) in to a new ones and which players are central promoting those transformations in the sets.
6. The dynamics of these centralities can be assessed by hypernetworks at different scales or levels of analysis, like: players individually, specific sets (groups) and sets of sets. Some studies used this hypernetworks multilevel approach considering the complex systems link between micro-meso-macro levels (Ramos et al., 2017a, 2017b; Ramos, 2019; Ribeiro et al., 2019).

► Example 5

Previous research revealed statistical significance between playing positions in successful and unsuccessful plays in football regarding flow centrality and flow betweenness (Korte et al., 2019). Defenders and defensive midfielders are functioning as bridging players in 70–75% of all plays they are involved in, the shares for goalkeeper and forwards are only 40–50%. ◀

► Example 6

The match-level metrics measure the share in a team total passing while the play-level metric evaluates the prevalence in plays across a match. For instance, flow betweenness detects how often a player is actually in-between two other players during a play and is in fact acting as an intermediary player. ◀

Study Box


The hypernetworks approach to PA in team ball sports has been tackling centralities either in cooperation and competition interactions (Ramos et al., 2017a, 2017b; Ramos, 2019; Ribeiro et al., 2019, 2020). The promising results were mainly based on distance interactions between players and allowed to identify some centralities:

1. The most common sets of players formed by proximity to each other were 1 vs. 1 (25%), followed by 1 vs. 2 (10.31%), 2 vs. 1 (8.8%) and 2 vs. 2 (6.81%);
2. The sets positioning interactions (e.g. through heatmaps) tends to be a reflex of the strategy (design) for those players roles on the match;
3. The synchronization processes between team players and opponents that emerged in the matches;
4. The dynamical changes in the sets were promoted by changes on the players running lines velocity.

► Example 7

The quantification of network centrality within a team or between teams provides an assessment of each player's mechanism of contribution. More specifically, the Eigenvector centrality would reflect the specific network structures of one's neighbor vertices. Sasaki and colleagues did not used betweenness centrality but rather the eigenvector centrality, which accordingly to the authors reflected the vertex centrality strongly (Sasaki et al., 2017). ◀

? Questions for the Students

1. How can we identify the playmakers during the course of invasion team sports match?
2. What are the type of interactions studied using network metrics?
3. Consider the network representing the passes between players in the Germany-Ghana match (WorldCup2014). Compute and/or identify:
 - (a) The order of the network.
 - (b) The team that has a denser pass network
 - (c) The most central player in each team (identify the criteria used).
 - (d) Which pair of players (in each team) have a more and less reciprocal passing relation.
 - (e) In the previous items (3a. to 3d.) I was performing: _____
4. Considering  Fig. 19.2 how do you describe the players' relations based on proximity? How would this change if player r_2 moved to support player r_1 ?

References

- Araujo, D., Davids, K., Bennett, S., Button, C., & Chapman, G. (2004). Emergence of sport skills under constraints. In *Skill acquisition in sport* (pp. 409–434). Routledge, Taylor & Francis e-Library.
- Balkundi, P., & Kilduff, M. (2006). The ties that lead: A social network approach to leadership. *The Leadership Quarterly*, 17(4), 419–439.
- Bar-Yam, Y. (1997). *Dynamics of complex systems*. Addison-Wesley.
- Bar-Yam, Y. (2003). *Complex systems and sports: Complex systems insights to building effective teams*. NECSI.
- Bar-Yam, Y. (2004). *Making things work: Solving complex problems in a complex world*. Knowledge Industry.
- Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C. I., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., & Zanin, M. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, 544(1), 1–122.
- Brandes, U., Freeman, L. C., & Wagner, D. (2012). Social networks. In R. Tamassia (Ed.), *Handbook of graph drawing and visualization*. CRC Press.
- Brandes, U., Robins, G., McCranie, A., & Wasserman, S. (2013). What is network science? *Network Science*, 1(1), 1–15. <https://doi.org/10.1017/nws.2013.2>
- Clemente, F. M., Couceiro, M. S., Martins, F. M. L., & Mendes, R. S. (2014). Using network metrics to investigate football team players' connections: A pilot study. *Motriz: Revista de Educação Física*, 20(3), 262–271.
- Clemente, F. M., Couceiro, M.S., Martins, F.M.L., & Mendes, R.S. (2015a). Using network metrics in soccer: A macro-analysis. *Journal of Human Kinetics*, 45(1), 123–134.
- Clemente, F. M., Martins, F. M. L., Kalamaras, D., Wong, P. D., & Mendes, R. S. (2015b). General network analysis of national soccer teams in FIFA World Cup 2014. *International Journal of Performance Analysis in Sport*, 15(1), 80–96.
- Clemente, F. M., Martins, F. M. L., Wong, P. D., Kalamaras, D., & Mendes, R. S. (2015c). Midfielder as the prominent participant in the building attack: A network analysis of national teams in FIFA World Cup 2014. *International Journal of Performance Analysis in Sport*, 15(2), 704–722.
- Criado, R., Romance, M., & Vela-Pérez, M. (2010). Hyperstructures, a new approach to complex systems. *International Journal of Bifurcation and Chaos*, 20(03), 877–883.
- Davids, K., Glazier, P., Araújo, D., & Bartlett, R. (2003). Movement systems as dynamical systems. *Sports Medicine*, 33(4), 245–260.
- Davids, K., Araújo, D., & Shuttleworth, R. (2005). Applications of dynamical systems theory to football. In *Science and Football V* (pp. 537–550).
- Duarte, R., Araújo, D., Correia, V., & Davids, K. (2012). Sports teams as superorganisms: Implications of sociobiological models of behaviour for research and practice in team sports performance analysis. *Sports Medicine*, 42(8), 633–642.
- Duch, J., Waizman, J. S., & Amaral, L. A. N. (2010). Quantifying the performance of individual players in a team activity. *PLoS One*, 5(6), e10937.
- Dutt-Mazumder, A., Button, C., Robins, A., & Bartlett, R. (2011). Neural network modelling and dynamical system theory. *Sports Medicine*, 41(12), 1003–1017.
- Fewell, J. H., Armbruster, D., Ingraham, J., Petersen, A., & Waters, J. S. (2012). Basketball teams as strategic networks. *PLoS One*, 7(11), e47445.
- Freeman, L. C. (2000). Visualizing social networks. *Journal of Social Structure*, 1(1), 4.
- Grund, T. U. (2012). Network structure and team performance: The case of English premier league soccer teams. *Social Networks*, 34(4), 682–690.
- Guillaume, J.-L., & Latapy, M. (2006). Bipartite graphs as models of complex networks. *Physica A: Statistical Mechanics and its Applications*, 371(2), 795–813.
- Johnson, J. (2006). Hypernetworks for reconstructing the dynamics of multilevel systems.
- Johnson, J. (2008). Multidimensional events in multilevel systems. In D. In Albeverio, P. G. Andrey, & A. Vancheri (Eds.), *The dynamics of complex urban systems: An interdisciplinary approach*. S (pp. 311–334). Heidelberg, Physica-Verlag HD.

- Johnson, J. (2013). *Hypernetworks in the science of complex systems*. Imperial College Press London.
- Johnson, J. H. (2016). Hypernetworks: Multidimensional relationships in multilevel systems. *The European Physical Journal Special Topics*, 225(6), 1037–1052.
- Katz, N., Lazer, D., Arrow, H., & Contractor, N. (2004). Network theory and small groups. *Small Group Research*, 35(3), 307–332.
- Koh, S., Yamamoto, T., Murakami, J., & Ueno, Y. (2013). Defence performance analysis of Rugby Union in Rugby World Cup 2011: Network analysis of the turnover contributors. In *Performance Analysis of Sport IX* (pp. 120–125). Routledge.
- Korte, F., Link, D., Groll, J., & Lames, M. (2019). Play-by-play network analysis in football. *Frontiers in Psychology*, 10, 1738.
- Moody, J., McFarland, D., & Bender-deMoll, S. (2005). Dynamic network visualization. *American Journal of Sociology*, 110(4), 1206–1241.
- Newman, M. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27(1), 39–54.
- Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.
- Passos, P., Silva, R. A. E., Gomez-Jordana, L., & Davids, K. (2020). Developing a two-dimensional landscape model of opportunities for penetrative passing in association football—Stage I. *Journal of Sports Sciences*, 38(21), 2407–2414.
- Passos, P., Davids, K., Araújo, D., Paz, N., Minguéns, J., & Mendes, J. (2011). Networks as a novel tool for studying team ball sports as complex social systems. *Journal of Science and Medicine in Sport*, 14(2), 170–176.
- Pina, T. J., Paulo, A., & Araújo, D. (2017). Network characteristics of successful performance in association football. A study on the UEFA champions league. *Frontiers in Psychology*, 8, 1173. <https://doi.org/10.3389/fpsyg.2017.01173>
- Ramos, J., Lopes, R. J., Marques, P., & Araújo, D. (2017a). Hypernetworks reveal compound variables that capture cooperative and competitive interactions in a soccer match. *Frontiers in Psychology*, 8, 1379.
- Ramos, J., Lopes, R. J., Marques, P., & Araújo, D. (2017b). Hypernetworks: Capturing the multilayers of cooperative and competitive interactions in soccer. *International Congress Complex Systems in Sport, Frontiers*.
- Ramos, J., Lopes, R. J., & Araújo, D. (2018). What's next in complex networks? Capturing the concept of attacking play in invasive team sports. *Sports Medicine*, 48(1), 17–28.
- Ramos, J. (2019). *Complex networks analysis in team sports performance: Multilevel Hypernetworks approach to soccer matches*. ISCTE-Instituto Universitário de Lisboa (Portugal).
- Reilly, T., Cabri, J., & Araújo, D. (2005). Applications of dynamical systems theory to football. In *Science and Football V* (pp. 570–572). Routledge.
- Ribeiro, J., Davids, K., Araújo, D., Silva, P., Ramos, J., Lopes, R., & Garganta, J. (2019). The role of hypernetworks as a multilevel methodology for modelling and understanding dynamics of team sports performance. *Sports Medicine*, 49, 1337–1344.
- Ribeiro, J., Lopes, R., Silva, P., Araújo, D., Barreira, D., Davids, K., Ramos, J., Maia, J., & Garganta, J. (2020). A multilevel hypernetworks approach to capture meso-level synchronisation processes in football. *Journal of Sports Sciences*, 38(5), 494–502.
- Sasaki, K., Yamamoto, T., Miyao, M., Katsuta, T., & Kono, I. (2017). Network centrality analysis to determine the tactical leader of a sports team. *International Journal of Performance Analysis in Sport*, 17(6), 822–831.
- Vilar, L., Araújo, D., Davids, K., & Button, C. (2012). The role of ecological dynamics in analysing performance in team sports. *Sports Medicine*, 42(1), 1–10.



Artificial Neural Networks

Markus Tilp

Contents

20.1 Example Sport – 170

20.2 Background – 171

20.3 Applications – 172

References – 176

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>.

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Artificial neural networks are inspired by the biological nervous system in their structure and function.
- They consist of so-called neurons, which are arranged in several layers. Connections exist between the neurons of the individual layers, which transmit signals between the neurons and thereby excite or inhibit them.
- Artificial neural networks convert input information into an output signal without the need to know the specific context.
- There are different ways in which artificial neural networks work: e.g., one distinguishes between networks that learn contexts supervised or unsupervised (i.e., completely independently).
- It has been shown that artificial neural networks can be used in sports, e.g., to identify patterns or to make predictions.
- It is expected that Artificial Neural Networks will be increasingly and successfully used in various fields of sports science in the future.

20.1 Example Sport

The biological nervous system adapts when a movement is learned or when a tactical situation is repeatedly observed. In the first case, if a complex movement is repeated many times by activating muscles which are controlled by a great number of nerve cells, the nervous system remembers the interaction of the corresponding nerve cells (Des Marées, 2003). During this process, the nerve cells receive information via the so-called dendrites and, if a threshold value is reached that activates the nerve cell, they in turn pass on information to the next cell via their axon. These structures and the synapses, i.e., the connections between the cells, strengthen if they are often used or atrophy if they are not used. After a training phase, the movement thus becomes increasingly smoother and more economical. In the second case, if certain game situations occur again and again, the sensory nervous

system will always perceive them similarly. Over time, the nervous system learns to identify the patterns that occur and can recognize the situations, e.g., a feint in a sports game.

Artificial neural networks work in a similar way. Their neurons receive information from other neurons and pass it on. While the biological synapses adapt over time to improve the signal transmission and thereby excite or inhibit it, artificial neural networks adapt the computational rules between two neurons. Similar to the biological process, the phase of this adaptation is also called training. If a neural network has trained/learned sufficiently by using input data, it can, for example, recognize patterns of play in a sports game (cf. Grunz et al., 2012; Memmert & Perl, 2009a, 2009b; Perl et al., 2013).

20.2 Background

There are various types of network structures, of which the simplest will be illustrated here. If, at one hand, information is always passed on in one direction only (from an input layer towards an output layer), this is referred to as a feedforward network. If, on the other hand, the network can fall back on earlier computations by allowing feedback, it is called a feedback network (backpropagation).

The basic structure of a feedforward neural network consists of an input layer, which in turn consists of several neurons, one or more hidden layers, and an output layer. The individual neurons of one layer are connected to all neurons of the next layer (see Fig. 20.1). Conversely, each neuron receives information from all neurons in the previous layer. The connections determine to what extent a value of one neuron is passed on to the next. This is called the weight of the connection. The weight of the connection of e.g., neuron 2 and neuron 5 from the next layer in Fig. 20.1 is denoted by w_{25} . The input value of neurons results from the weighted values of all upstream neurons. These values can be combined by linear or nonlinear calculation rules. This calculation rule is called propagation function and is in the simplest case the sum of all weighted values of the incoming neuron values (see calculation in Example 1).

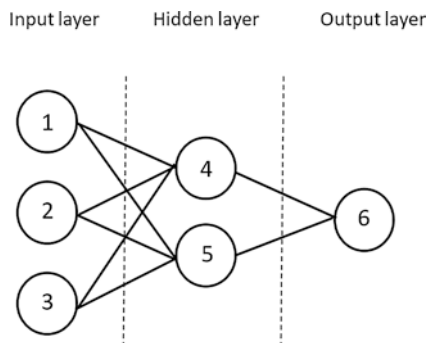


Fig. 20.1 Basic structure of a neural network with input layer (neurons 1–3), hidden layer (neurons 4 and 5), and output layer (neuron 6)

The input value of a neuron from the propagation function is now further processed via a so-called activation function (within the receiving neuron). Similar to the way a neuron becomes active in nature only when a certain membrane potential is exceeded, there are computational rules that calculate the activity value of the neuron from the input value. There are different types of activation functions (e.g., a step function, which produces a constant value above a certain input threshold). The resulting activity value can then be forwarded to the next layer or, in the case of an output layer neuron, represents the output value.

The weights between the neurons are variable and can adapt to the data. The phase of these adaptations, in which a set of input data is provided to the neural network, is called the training or learning phase. During this phase, the weights between the neurons are changed according to predefined learning rules. A distinction is made between supervised and unsupervised learning. In supervised learning, the network is provided with both input data and the corresponding output data, just as a teacher tells the student the task and the solution. During the learning process, the network compares the calculated and the given output values and continuously adjusts the weights based on the differences until the differences are below a given threshold. The way the differences are calculated are specified in a so-called error function. Network types for supervised learning are e.g., perceptron, multi-layer perceptron (MLP), or radial basis function (RBF) networks. During unsupervised learning, only input data is provided. The network then tries to map the input data to neighboring neurons based on their similarity. Hence, similar input data are assigned to similar output data after the training phase. One network type for unsupervised learning is, for example, a Kohonen feature map (KFM). Thus, the choice of network type depends on the task that it is supposed to solve. For exact descriptions of the different network types, learning rules, propagation, activation, and error functions, the reader is referred to further literature (e.g., Backhaus et al., 2006; Sanderson, 2017).

Definition

Artificial neural networks are information-processing systems that are modeled based on the networking of nerves in living organisms. They consist of so-called neurons, which are arranged in individual layers. The connections between the neurons link the input information with output information. Artificial neural networks are a sub-area of artificial intelligence and are used for classification, forecasting, and optimization tasks.

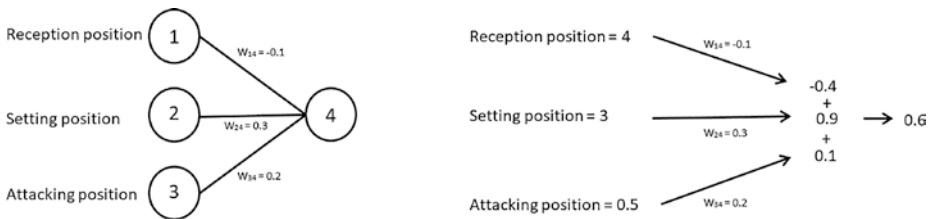
20.3 Applications

► Calculation Example 1

Here, we create an artificial neural network for beach volleyball to predict whether an opponent will play a hard attacking spike or a targeted shot. We assume that the playing positions during the reception, setting, and attacking affect this decision. Therefore, the

distance of the attacker's receiving position from the net (in m), the distance of the setting position from the net (in m), and the distance of the attacking position from the net (in m) serve as the net input information. Since this is a supervised learning process, the network is also informed about the type of attack (0 = spike, 1 = shot) as output information. In this case, the artificial neural network consists of an input layer with three neurons (distances between 1—the reception position, 2—the set position, and 3—the attack position to the volleyball net, respectively) and an output layer with one neuron (type of attack). The weights between the neurons are randomly selected at the beginning and adjusted during the learning process. As a learning rule, the weights could, for example, always be increased or decreased by 5% of the input values of the current situation (learning rate), depending on whether the result was too low or too high. In contrast, if the calculation leads to the correct result, the weights remain unchanged (■ Fig. 20.2).

■ Table 20.1 shows an example data set with data collected from a game. The calculations with this data should explain how an artificial neural network works. If we take the data of the first row from the example data set (■ Table 20.1), the input data and the corresponding weights $w_{14} = -0.1$, $w_{24} = 0.3$, and $w_{34} = 0.2$ result in a value of 0.6 ($4 \times (-0.1) + 3 \times 0.3 + 0.5 \times 0.2$) (see ■ Table 20.2). This value is then further processed by an activation function. In the example, a so-called threshold function would be appropriate as activation function, e.g., taking the value 0 (spike) for values < 0.5 and the value 1 (shot) for values ≥ 0.5 . Applying the activation function would result in the value 1 (Shot) in our calculations since 0.6 is greater than 0.5. However, since the player played an attack (value 0) in the real situation (see ■ Table 20.1), the weights would be adjusted before calculating the second game situation. In this case, the weights



■ Fig. 20.2 Calculation example 1: Left: General structure of the type of attack prediction network. Right: Calculation of the input value of the output neuron based on the input data and the weights w_{ij} in situation 1 from the example data set

■ Table 20.1 Example data set

| | Distance reception | Distance setting | Distance attack | Attack behavior |
|-------------|--------------------|------------------|-----------------|-----------------|
| Situation 1 | 4 | 3 | 0.5 | 0. Spike |
| Situation 2 | 6 | 1 | 1 | 1. Shot |
| Situation 3 | 4.5 | 2 | 0.5 | 0. Spike |

Table 20.2 Input data (e_i), weights (w_{ij}), weighted output values, and sums (=input values) from the example data set. The weights always increase or decrease by 5% of the input data

| | Recep-
tion =
e_1 | Setting
= e_2 | Attack
= e_3 | w_{14} | w_{24} | w_{34} | $e_1 w_{14}$ | $e_2 w_{24}$ | $e_3 w_{34}$ | Sum |
|--------|---------------------------|--------------------|-------------------|----------|----------|----------|--------------|--------------|--------------|--------|
| Sit. 1 | 4 | 3 | 0.5 | -0.1 | 0.3 | 0.2 | -0.4 | 0.9 | 0.1 | 0.6 |
| Sit. 2 | 6 | 1 | 1 | -0.3 | 0.15 | 0.175 | -1.8 | 0.15 | 0.175 | -1.475 |
| Sit. 3 | 4.5 | 2 | 0.5 | 0 | 0.2 | 0.225 | 0 | 0.4 | 0.1125 | 0.5125 |

would be reduced by 5% of the input values from situation 1. The new weights would be $w_{14} = -0.1 - 0.05 \times 4 = -0.3$, $w_{24} = 0.3 - 0.05 \times 3 = 0.15$, and $w_{34} = 0.2 - 0.05 \times 0.5 = 0.175$. Now, using the new input data from situation 2 (reception position = 6 m, setting position = 1 m, attacking position = 1 m) and the new weights, the new input value is calculated. This results in a value of $-1.475 (= 6 \times (-0.3) + 1 \times 0.15 + 1 \times 0.175)$. Since this value is < 0.5 , the activation function would yield the value 0 (spike). This value again does not match the observed behavior (shot), so the weights are again adjusted. This process is repeated with an amount of input data as large as possible until the predictions reach a given sufficient accuracy. Then the network has learned the context and can predict whether the player will spike the ball or play a shot based on data from a (new) input data set (reception position, setting position, attacking position). ◀

► Example 2 Pattern Recognition

In handball, teams are interested in tactical moves that are supposed to result in favorable goal-throwing situations. During these moves, the players and the ball cover pre-arranged paths of travel and passing. Now it is e.g., of interest to identify such moves of the opposing team to better anticipate their actions. The identification of such patterns is a typical task for artificial neural networks. Based on positional data of passing positions from handball games, Schrapf and Tilp (2013) used an artificial neural network (Kohonen feature map) to identify teams' moves. Positional data from 612 action sequences, each consisting of the position of the throw and the preceding five passing stations, were used. To achieve a larger data set of 3060 action sequences, the data was amplified by adding random noise. With this data set, the neural network was then able to identify 42 different types of attacking moves. By inspecting the corresponding video sequences, experts confirmed that the attacking moves identified by the neural network corresponded to actual game moves. Interestingly, 49% of all attacks consisted of only 8 types of attack, i.e., only a small selection of attacks was preferentially used. For validation, the coordinates of the real data could subsequently be compared with the attacks identified by the network. The average deviation was only 1.2 m and hence, sufficiently small. An advantage of using a neural network in comparison to classical statistical analyses is, that in addition to the throwing position, the preceding passing positions are also included in the analyses. This makes it possible to analyze also the emergence of throwing situations. ◀

► Example 3 Situation Prediction/Forecast

Ground reaction forces during locomotion provide information about the load and performance during walking and running. However, measuring ground reaction forces is very complex, expensive, and usually only possible in laboratories. Komaris et al. (2019) used Artificial Neural Networks (supervised, feed-forward with input, output, and a hidden layer) to calculate ground reaction forces of the three spatial directions (x, y, z) at three different speeds (2.5 m/s, 3.5 m/s, 4.5 m/s) via motion data. An artificial neural network was used for each spatial direction. The used data were the acceleration data of the lower legs from 3D motion analyses (= input data) as well as the corresponding force curves (= output data) from a treadmill with force plates of 28 professional runners. The acceleration and force curve trajectories were scaled so that each data set contained 100 data points (1–100% of the motion cycle). Accordingly, the input and output layers also consisted of 100 neurons each. The hidden layer consisted of 10 neurons. In the training phase, the data sets of 16 randomly selected subjects were used, and in the validation and testing phase, the data sets of 6 subjects were used. The differences between the calculated and true ground reaction forces were independent of velocities and sufficiently small so that the method can be recommended to estimate ground reaction forces from lower leg accelerations. These accelerations can be collected relatively easily, e.g., via inertial measurement units (a combination of several inertial sensors, e.g., accelerometers and gyroscopes). ◀

? Questions for the Students

1. What are the different layers of an artificial neural network called?
2. How is the value of a neuron calculated?
3. How does a neural network learn during the training phase?
4. Give two concrete examples of how it can be used in sports.

Study Box

Especially in sports games, anticipation (predicting game actions) is an important factor for success. It is known that experts in a sports game can recognize the actions of their opponents early, e.g., based on observations of the opponent's position or posture. Schrapf et al. (2022) used an artificial neural network (multi-layer perceptron) to predict attack positions in indoor volleyball. In indoor volleyball, block players need to know the position of the attack as early as possible to get into a good position for the block. The researchers trained their neural network with the positions of the receiving player at the time of reception as well as the setter at the time of reception and setting, the trajectory of the reception (time), the type of setting (lower or upper setting), and the type of movement during the setting (standing, jumping, running, diving). In addition, during the supervised training phase the network was also provided with the attacking position (positions 2, 3, 4, and back court) and the setting time (<0.8 s, 0.8–1.2 s, >1.2 s). The predictions of the neural network were then compared with the predictions of experts (coaches at the national team level). They had to estimate the attacking position and passing time from a video that was stopped shortly before the setting. Although the neural network had

only limited information of input data compared to the experts who saw all players in the video, the predictions of the attacking position were about equally good (68.1% vs. 65.3% correct predictions) and for the setting time even significantly better (79.2% vs. 64.6% correct predictions) than those of the experts. The result shows that neural networks are capable of predicting actions in sports games with similar success rates as experts. It has to be noted that the results could probably be even improved by varying and optimizing the input data. By pointing out the essential prediction parameters identified by the neural network, these results could then be used in the context of perception training. A corresponding anticipation training is also possible for athletes in other sports such as handball (Hassan et al., 2017).

References

- Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2006). Neuronale Netze. In *Multivariate Analysemethoden* (11. Auflage, pp 750–806). Springer, Berlin.
- Des Marées, H. (2003). *Sportphysiologie* (9. Auflage ed.). Sportverlag Strauß.
- Grunz, A., Memmert, D., & Perl, J. (2012). Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human Movement Science, 31*, 334–343.
- Hassan, A., Schrapf, N., & Tilp, M. (2017). The prediction of action positions in team handball by non-linear hybrid neural networks. *International Journal of Performance Analysis in Sport, 17*, 293–302.
- Komaris, D. S., Pérez-Valero, E., Jordan, L., Barton, J., Hennessy, L., O'Flynn, B., & Tedesco, S. (2019). Predicting three-dimensional ground reaction forces in running by using artificial neural networks and lower body kinematics. *IEEE Access, 7*, 156779–156786.
- Memmert, D., & Perl, J. (2009a). Analysis and simulation of creativity learning by means of artificial neural networks. *Human Movement Science, 28*, 263–282.
- Memmert, D., & Perl, J. (2009b). Game creativity analysis by means of neural networks. *Journal of Sport Science, 27*, 139–149.
- Perl, J., Grunz, A., & Memmert, D. (2013). Tactics in soccer: An advanced approach. *International Journal of Computer Science in Sport, 12*, 33–44.
- Sanderson, G. (2017, August 1). Neural networks. [Video]. YouTube. https://www.youtube.com/watch?v=aircAruvnKk&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi.
- Schrapf, N., Hassan, A., Wiesmeyr, S., & Tilp, M. (2022). An artificial neural network predicts setter's setting behavior in volleyball similar or better than experts. *IFAC-PaperOnLine* (55–20, pp. 612–617).
- Schrapf, N., & Tilp, M. (2013). Action sequence analysis in team handball. *Journal of Human Sport and Exercise, 8*(3), 615–621.



Deep Neural Networks

Dominik Raabe

Contents

21.1 Example Sport – 178

21.2 Background – 179

21.3 Applications – 180

References – 183

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>.

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com.

Key Messages

- Deep neural networks are neural networks with multiple layers and potentially millions of parameters and potentially billions of links.
- They are successfully used for particularly complex learning problems in research and industrial applications in diverse fields such as image recognition or language processing.
- The ability to master complex learning problems is usually accompanied by high training data requirements, intensive computation times, and low transparency of the trained model.
- In sports, there are countless applications for deep neural networks, from data acquisition to semantic event extraction to analysis of complex tactical patterns.

21.1 Example Sport

Deep neural networks are an extension and development of classical neural networks (see ► Chap. 20) in the field of artificial intelligence. While neural networks in the early years of their development often consisted of only a few neurons and at most one hidden layer, modern "deep" neural networks can consist of several hundred million neurons, arranged in a multitude of hidden layers and billions of connections between the neurons. Accordingly, the difference between these networks and classical networks is the number of layers and neurons. Sometimes, however, the term "deep learning" is nowadays used excessively for other methods in order to emphasize the particular complexity of a model. Closely related to the size of these networks, however, is the leap in performance, as deep neural networks are now clearly ahead of their smaller ancestors on numerous problems on standardized datasets (so-called benchmark datasets).

Due to this performance, deep neural networks are extremely popular nowadays and are used for application or research problems in a wide variety of domains such as image recognition and processing (so-called computer vision), speech rec-

ognition, text processing, medicine, bioinformatics, climate research, but also in sports. For example, it is only the advances in computer vision that have favored the development of optical (i.e., camera-based) tracking methods for acquiring positional data with acceptable accuracy. With these methods, the players and their positions are detected and extracted from the camera image with the help of deep neural networks. The very latest methods can even extract the positions from a television broadcast or estimate the exact silhouette of the players (see ► Chap. 4). Also, in the area of sports data analysis, there are more and more methods that can be classified as deep learning.

21.2 Background

Although some historical precursors of modern deep networks were developed at the end of the twentieth century, the breakthrough of these methods can be dated to 2012. In that year, deep neural networks were able to rival conventional methods in speech recognition for the first time (Hinton et al., 2012). In addition, several new records were set in image recognition on common benchmark datasets, thanks to a combination of network architecture, size, as well as computational power (Ciresan et al., 2012; Krizhevsky et al., 2017). Here, so-called Convolutional Neural Networks (CNNs) were used, which are modeled in a rough analogy to the human visual cortex in their structure and vividly illustrate the characteristics of deep learning.

These CNN architectures consist of several layers, which have a modular structure and are connected sequentially. Data—in this case images encoded as matrices of pixel color values—are fed into the network via an input layer and then passed step by step through the individual layers, each layer having its own task. Accordingly, earlier layers are meant to detect basic properties (so-called features) of the image, such as contrasts or individual color areas, while later layers are meant to detect more complex features, such as initially lines and shapes up to body parts or objects. This logic follows the assumption that “higher-level” concepts such as a face are composed of “lower-level” concepts such as the eyes and nose, which in turn are composed of lines or dots. This hierarchical structure of concepts is one of the foundations of large network architectures (LeCun et al., 2015).

Accordingly, the complexity of modern network architectures has multiplied dramatically in recent years. Although they still largely consist of neurons, activation functions and propagation functions, their connections are becoming increasingly advanced. Fully connected layers, i.e., connecting each pair of neurons between two successive layers, that characterized earlier neural networks is now only one of countless modules. Instead, there are, for example, modules that feedback neuron output to previous network layers (so-called recurrent neural networks), or randomly delete individual data points within the network to enable more robust training (so-called dropout). The search for new modules as well as the connection of these modules to architectures is an essential part of the current research.

However, deep neural networks also have some drawbacks associated with their sheer size. These networks consist of hundreds to billions of “trainable” parameters—e.g. weights of the individual propagation functions—which are adjusted during the learning process. This causes both practical and theoretical problems. To adjust the individual parameters, i.e., to “learn” the solution, these methods typically require a lot of training data (in supervised learning), long computation times, and associated resources such as hardware or energy. One of the largest networks constructed to date, the GPT-3 in the text processing domain, boasts 175 billion parameters (Brown et al., 2020). The training of this model used a total of 45 terabytes of compressed text material (about 400 billion characters) and ran on a specifically constructed supercomputer with 10,000 processing units for about a month. The estimated cost is in the range of several million US dollars, and the runtime, extrapolated to a single computing unit, is several hundred years. Although this is an extreme example, it shows that feasibility is limited both for problems with smaller data sets (some of which have to be created by hand) and by financial constraints.

Furthermore, due to the complexity of these models, it is not possible to comprehend the final predictions of a trained network. Due to the high number of parameters, these models are usually a “black box” with opaque inner computational mechanisms. Why a network makes a certain prediction remains, for the most part, its secret. Apart from that, from a scientific perspective, simple and transparent models are preferable to more complex and non-transparent models according to Occam’s razor. Because of these disadvantages, research today is intensively focused on models that learn on the basis of very little training data (e.g. so-called few-shot learning) and that deliver predictions that are comprehensible to humans (so-called explainable artificial intelligence, abbreviated “xAI”).

Definition

Deep neural networks and deep learning refers to a group of machine learning methods, and in particular neural networks, that are characterized by their complexity and size. Such large network architectures consist of several submodules with different structure and functionality, which are modularly linked and coupled with each other. This approach follows the assumption that the concept to be learned in the real world is composed of several hierarchically structured sub-concepts. In a deep learning method, these are to be learned sequentially up to the final concept through the different network layers—hence the term “deep” learning.

21.3 Applications

A classic example for the use of deep neural networks in the context of sports is the generation of position data via so-called optical, i.e., camera-based tracking systems. These systems are characterized by their non-invasiveness, since the players do not have to wear additional sensors on their bodies. With the help of neural

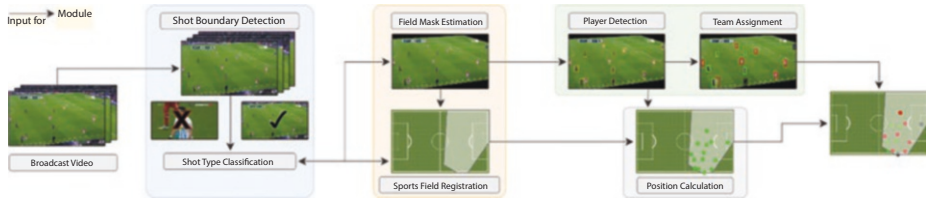


Fig. 21.1 Schematic structure of an exemplary processing pipeline for extracting position data from a television broadcast. (Adopted from Theiner et al. 2022)

networks, the players can be detected on the camera image, and using up to 20 cameras with different viewing angles, the exact positions can be inferred. While these cameras are usually fixed and calibrated in the stadium infrastructure, the technology is now so advanced that a dynamic camera image (from a TV broadcast, for example) alone can be used to extract player positions. However, this requires several steps besides the identification of the players to obtain accurate positions. Theiner et al. (2022) show an example of an entire pipeline that can solve this task (see [Fig. 21.1](#)). This pipeline consists of steps to identify the individual scene cuts, select all sequences that are filmed from the main camera, recognize the playing field and estimate the field calibration, recognize the players, extract the positions, and assign identities. In almost all of these steps, deep learning methods are used and show higher precision than conventional methods.

The study by Wagenaar et al. (2017), on the other hand, uses already processed position data and addresses the question of whether goal chances can be predicted using deep neural networks (the CNNs described above). For this purpose, several network architectures with increasing complexity were used in an experiment. From positional data of soccer matches, the authors extracted short sequences of 10 s in length, which either ended in a goal scoring opportunity or not. Based on this dataset, a binary classification problem was formulated and the task for the networks was to decide, given a sequence shown, whether a goal was scored—or not. Since, as described, CNNs originate from image processing, the raw position data were first transformed into 256×256 pixel images that schematically represent the respective game scene in a two-dimensional pixel graphic (see [Chap. 23](#)). This representation of the raw data was used as input data for the different networks. A comparatively small and simple CNN, an extremely performant (at that time) GoogLeNet, and a K-nearest neighbor approach were used as baselines. Compared to the baseline with an average classification accuracy of 57.3%, the deep neural networks were able to achieve significantly better results: the best GoogLeNet variant achieved an average accuracy of 67.1%.

? Questions for the Students

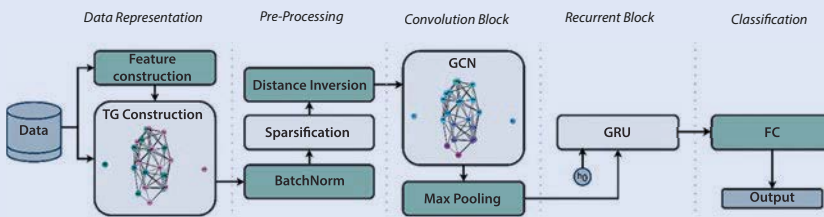
1. What distinguishes deep neural networks from “classical” neural networks?
2. What are the disadvantages of deep neural networks and what causes them?
3. What are three exemplary classes of deep neural networks?
4. What characterizes the basic structure of deep neural networks?

Study Box

The study described above is one of the first to attempt to predict game tactical events such as goal completions based on positional data and using deep learning (Wagenaar et al., 2017). However, the question arises to what extent a representation of position data in pixel graphics is appropriate for the specific problem. More generally, the question arises which data representation and which network modules or architectures are particularly suitable for processing position data in deep learning models. The choice of these has a decisive influence on the performance in sports-specific learning problems, as well as on the associated model complexity and the required sample sizes in the learning process. In a similar study, Raabe et al. (2022) therefore propose a graph-based data representation and a matching deep neural network architecture consisting of several submodules (see Fig. 21.2). Here, the raw positions of the players are represented as nodes in a graph, which are linked by their respective interactions. This representation is intended to exploit domain-specific characteristics of sports and thus improve the performance of the network. Furthermore, graphs have interesting mathematical properties (so-called invariances) that reduce the need for training data. Graph-based neural networks are a recent development in the field of deep learning, which enables the processing of non-Euclidean input data.

In an experimental comparison, this neural network architecture was then tested against other architectures. As a learning problem, a binary classification task was constructed in which short sequences of positional data were to be evaluated with respect to the question whether the sequences resulted in a ball win for the defending team—or not. For comparison, a simple baseline and four different data representations were used. One was a pure feature-based approach, in which numerous common key performance indicators (KPIs) were calculated from the raw positional data, which were subsequently used as predictors in a logistic regression. Furthermore, deep neural networks based on the raw position data, i.e., a long vector with the player positions, as well as based on a representation as pixel images—as in the study by Wagenaar et al. (2017)—were utilized. Finally, the proposed graph-based architecture was also utilized.

The results of this comparison are summarized in the following table:



■ Fig. 21.2 The graph-based deep neural network architecture used from Raabe et al. (2022)

| Model | Data presentations | Number of parameters | Classification accuracy (%) |
|---------------------|--------------------|----------------------|-----------------------------|
| Baseline | – | 1 | 74.5 |
| Logistic regression | KPIs | 10 | 76.8 |
| SVGRU | Raw data | 499.970 | 71.6 |
| CNN | Pictures | 437.506 | 61.4 |
| GoogLeNet | Pictures | 9.936.038 | 80.4 |
| TGNet | Graphs | 109.212 | 80.5 |

In summary, both the most complex GoogLeNet with almost ten million parameters and the proposed graph-based model TGNet achieved by far the best classification results. However, the TGNet is much leaner, requires only a fraction of the parameters and the required training, inference, and adaptation times are much lower (Raabe et al., 2022). Thus, it can be concluded that the development and selection of appropriate deep neural network architectures is a key task in sports informatics in order to exploit the maximum potential of deep learning for sport-specific problems. However, to enable this development, other requirements are essential. In particular, for the further development of deep learning in the field of sports, suitable, publicly accessible benchmark data sets as well as transparent representations of used architectures are indispensable, which are significantly weaker in this domain than in those fields where deep learning has already achieved groundbreaking success (Raabe et al., 2022).

References

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodè, D. (2020). Language models are few-shot learners (arXiv:2005.14165). arXiv. <http://arxiv.org/abs/2005.14165>
- Ciresan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition* (pp 3642–3649). <https://doi.org/10.1109/CVPR.2012.6248110>.
- Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6), 82–97. <https://doi.org/10.1109/MSP.2012.2205597>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>

- Raabe, D., Biermann, H., Bassek, M., Wohlan, M., Komitova, R., Rein, R., Groot, T. K., & Memmert, D. (2022). Floodlight—A high-level, data-driven sports analytics framework. *Journal of Open Source Software*, 7(76), 4588. <https://doi.org/10.21105/joss.04588>
- Raabe, D., Nabben, R., & Memmert, D. (2022). Graph representations for the analysis of multi-agent spatiotemporal sports data. *Applied Intelligence*, 53, 3783–3803. <https://doi.org/10.1007/s10489-022-03631-z>
- Theiner, J., Gritz, W., Müller-Budack, E., Rein, R., Memmert, D., & Ewerth, R. (2022). Extraction of positional player data from broadcast soccer videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)* (pp 823–833).
- Wagenaar, M., Okafor, E., Frencken, W., & Wiering, M. A. (2017). Using deep convolutional neural networks to predict goal-scoring opportunities in soccer. In *International conference on pattern recognition applications and methods*. <https://doi.org/10.5220/0006194804480455>.



Convolutional Neural Networks

Yannick Rudolph and Ulf Brefeld

Contents

- 22.1 Example Sport – 186**
- 22.2 Background – 187**
- 22.3 Applications – 189**
- References – 192**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>.
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>.

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com.

Key Messages

- Convolutional neural networks (CNNs) are important machine learning models for images or data with comparable grid structure
- For example, in supervised learning, CNNs may extract features for classification and object detection
- CNNs are characterized by parameter sharing and by their properties regarding shifted inputs
- CNNs are useful for learning features of spatial situations given by positional data from team sports

22.1 Example Sport

In sports analyses, images, videos, or positional data (see ► Chaps. 4 and 6) can provide essential information on characteristics of interest, on good or bad actions of individual players or a team. Usually, this information is extracted from the respective sources by trained analysts. This process is time-consuming, however, because the analysts must sift through the entire material to discover relevant situations.

With the help of machine learning, the analysts' work can be partially automated by having the computer filter the data according to relevance. Additionally, by processing large amounts of data it is often possible to discover previously unknown statistical correlations.

Consider an *expected goals* (xG) model in soccer as a motivating example. The idea behind xG is to evaluate goal chances by probabilities. The probability indicates how often similar chances lead to a goal. The higher the xG, the more likely it is that a situation will lead to a goal. A suitable model would be, for example, a probabilistic classification by means of logistic regression, which could be learned with the help of observed events ("goal" or "no goal").

To do this, the goal opportunities must be processed and be available in a computer-readable representation that is suitable for the classification task. In machine learning, we extract features that describe the situations and that allow the computer to find a good representation of the concepts contained in the data. For our xG example, suitable features would include the distance and angle of the player to the goal, the number of opponents involved, an indicator of whether the goalie is in the goal, whether one or more players are blocking the shot path, etc. The list of candidate features is long, and their exact definition and implementation may require substantial effort and is often non-trivial.

Convolutional neural networks (CNNs) offer an alternative: First, goal opportunities must be converted into an image-like grid structure using positional data, in which, for example, each *pixel* in the grid corresponds to one square meter on the playing field and the positions of the players and the ball are marked accordingly. Goal opportunities transformed in this way can now be used directly as input to the CNN to learn the probability of a goal with a computer. The time-consuming extraction of features for the task can be omitted because the CNN itself generates the features that are best suited for the learning task. Notably, the CNN learns features and classification simultaneously (keyword: *end-to-end* learning). Especially when it is difficult to define relevant features manually, models that can be learned end-to-end usually achieve the best results. As with other end-to-end methods, these successes are accompanied by the disadvantage that neural networks are difficult to interpret and explain.

22.2 Background

The most common application of convolutional neural networks (CNNs) is in the context of machine learning on images. However, in principle, CNNs are suitable for all data with one- or two- and even higher-dimensional lattice structures. Other neural networks that would, in principle, also be suitable for extracting features from images or image-like data (fully-connected neural networks) are usually either too *shallow* or have very large numbers of parameters. In both cases, the models often cannot be learned efficiently.

Predecessors of modern CNNs (Fukushima, 1980) and early CNNs like the *LeNet* model (LeCun et al., 1998, see also ■ Fig. 22.2) were already able to classify handwritten digits based on images. In parallel, methods from classical computer vision determined visual features such as edges or gradients based on pixel values in images. For example, a common technique applied a scale invariant feature transformation (*SIFT*) and used the resulting features for a classification task that was a separate processing step.

Since the multiplication of data and computational power, CNNs efficiently extract features that lead to significantly better results than classical methods. For example, a CNN architecture trained on graphics processing units (GPUs) (*AlexNet*, Krizhevsky et al., 2012) won a classification competition on the important ImageNet dataset (Deng et al., 2009) in 2012 by a very large margin over classical methods.

Usually, several *convolutional layers* are combined in CNNs to process the data. In a convolutional layer, several *convolutional filters* with learnable parameters operate on the data. These operations can be thought of as moving the filters (in the two-dimensional case: a matrix with learnable entries) over the input and computing an output for each position of the filter by summing up the results of a point-wise multiplication with the underlying input. Because a convolutional filter is applied multiple times to an input, convolutional layers have comparatively fewer parameters than traditional fully-connected layers. In this context, we also speak of *parameter sharing*.

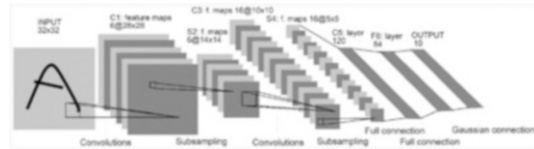
To visualize a filter operation see [Fig. 22.1](#) which shows an input, a convolutional filter, and the output resulting from the application of the filter. In the figure, input values used to calculate the upper left output value are highlighted (the calculation is given in the figure text). If the input has more than two dimensions (for example, color images have multiple color channels in addition to the height and width dimensions), the input and filter also have more dimensions. The output of a filter operation remains a matrix. If there are multiple convolutional filters in a layer, the respective outputs (also called *channels* or *feature maps*) are combined and form a multidimensional *tensor*.

The example calculation in [Fig. 22.1](#) is useful to illustrate a property of CNNs: If we shift the input one column to the right (padding it with zeros on the left, for example) and apply the same convolutional filter again, we will do the same calculation for the top right output value. That is: the output values will also shift to the right. This property of CNNs to generate an equally shifted output for shifted input is often described as *translation equivariance*—or as *translation invariance* in terms of the final representation. Graphically, translation invariance means that a CNN that has learned to classify balls in a particular region of an image can also recognize balls in any other region of an image. The network can thus *generalize*, i.e., learn valid relationships and successfully apply them to new data. For *true* translation equivariance or invariance, however, special conditions must be present (see also Kayhan & van Gemert, 2020).

Usually, CNN architectures use multiple convolutional layers, of which the first layer operates on the input and the subsequent layers operate on the respective outputs of earlier layers. Thereby, features in layers that are *deeper* (that is: closer to the final output) are usually influenced by a larger region of the input. Following the biology of the human eye, we also call this region the *receptive field*, since it is the area of the input that the feature can *see*. Because the receptive field grows with the depth of a layer, the further the input is propagated through the network, the more complex features CNNs can extract.

$$\begin{array}{|c|c|c|} \hline 1 & 7 & 5 \\ \hline 8 & 3 & 2 \\ \hline 9 & 4 & 6 \\ \hline \end{array} * \begin{array}{|c|c|} \hline 2 & 0 \\ \hline 1 & 3 \\ \hline \end{array} = \begin{array}{|c|c|} \hline 19 & 23 \\ \hline 37 & 28 \\ \hline \end{array}$$

Fig. 22.1 Exemplary representation of a filter operation; the sum of the point-wise multiplication (in color) is calculated as follows: $1 \cdot 2 + 7 \cdot 0 + 8 \cdot 1 + 3 \cdot 3 = 19$



■ **Fig. 22.2** The *LeNet-5* CNN architecture for classifying handwritten digits (LeCun et al., 1998)

■ Figure 22.2 illustrates a relatively simple CNN architecture by today's standards. In addition to convolutional operations, the figure also shows *subsampling* operations. In fact, many CNN architectures contain other layers in addition to convolutional layers, such as *pooling layers* for subsampling or so-called *residual layers* (He et al., 2016). Among other things, residual layers make it possible to successfully train very deep CNN architectures (that is: models with many layers). The scope of modern CNNs is not limited to classification or object detection: CNNs are also used in models for generating artificial images (e.g., Ramesh et al., 2021).

Definition

Convolutional Neural Networks (CNNs) are artificial neural networks that can efficiently learn representations and extract features, especially from two-dimensional data with a grid structure. The main feature of CNN architectures are convolutional layers, where point-wise products of filters and local neighborhoods in the input are aggregated into corresponding output values. One property that arises from this is that of translation invariance: the output of a CNN is (to some extent) not affected by shifts in the input.

22.3 Applications

► Example 1

Already early CNNs were able to successfully classify (handwritten) digits (see background). Gerke et al. (2015) showed that CNNs are also suitable for the automatic classification of numbers on player jerseys. To do so, the authors created a dataset of more than 8000 images of soccer players with legible jersey numbers and compared the classification using a CNN architecture to a baseline classification trained with features from classical image processing. The study concludes that a relatively flat CNN with three convolutional layers classifies numbers much better than the baseline method: while the baseline classified only 40% of numbers in unseen test images correctly, the CNN could classify numbers correctly in 83% of the test images. However, in Gerke et al. (2015), the detection of players in the images was still implemented using methods from classical image processing. ◀

► Example 2

CNNs can also be used for object detection in images. A study by Wei et al. (2016) illustrates how CNNs for object detection might be relevant for sports analytics: In the study, the authors propose so-called *convolutional pose machines*, which improve the detection of human postures in images—known as *pose prediction*—by using CNNs. Pose prediction is performed by detecting specific body parts that collectively describe the pose. The detection of body parts as proposed by Wei et al. (2016) is performed by the sequential application of CNN models. While the first CNN in this process operates only on the image data, subsequent CNNs also consider the output of previous models. The idea is, that information on relatively easy-to-recognize body parts (for example, head and shoulders) can be extracted using the first CNN models. Passing this information into subsequent CNNs can then enable the detection of more difficult-to-detect body parts (such as elbows). On a dataset of postures from eight sports (Johnson & Everingham, 2011), the authors were able to achieve the best result in terms of correctly localized body parts at the time. ◀

► Example 3

Regarding positional data in team sports, CNNs can be part of more complex models. An example of this is provided in a study by Fassmeyer et al. (2021), in which the authors investigate the automatic classification of game situations in soccer using positional data. The study considers positional data with 25 positions per second, where each position is converted into an image-like grid structure following the data transformation in Dick and Brefeld (2019). The models proposed by Fassmeyer et al. (2021) process both the spatial and temporal dimensions of the data: For each time step, features of the respective positions are extracted by a CNN and fed to a recurrent neural network (RNN). RNNs are models that are suited for sequential data. To make use of positions without a class label during training (keyword: semi-supervised learning), the models learn a representation of the data, from which the data itself can be reconstructed. For this purpose, the authors propose an *autoencoder* architecture. Empirically, Fassmeyer et al. (2021) report very good results for the classification of corners and edges based on the obtained representations. ◀

? Questions for the Students

1. Why are fully-connected neural networks usually not suited for machine learning on images or image-like data?
2. Under what conditions does processing two different images with the same CNN lead to the same representation?

Study Box

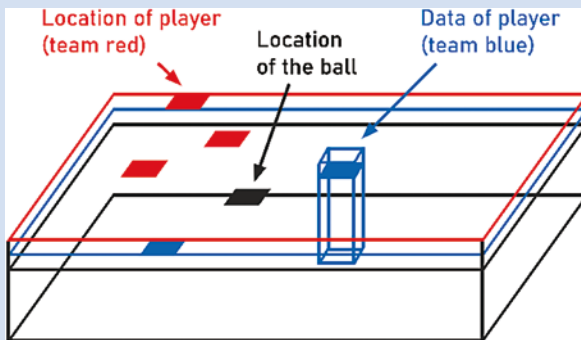
Positional data (see ► Chap. 6) is very much suitable for processing by CNNs. Dick and Brefeld (2019) for example propose a model that automatically evaluates positions in professional soccer using a CNN architecture which operates on a spatial representation of positional data. Specifically, the representation is based on the idea of dividing the soccer field into a two-dimensional grid in which the positional data

of players and ball are encoded. The data is thus transformed to an image-like grid structure. However, this data structure is not limited by the properties of images. Instead of working with three RGB color channels, the authors use separate channels to encode different information. Specifically, Dick and Brefeld (2019) propose nine channels, all initialized with zeros. See also the example in Fig. 22.3.

In the first channel all player locations of the first team are set to a value of “1”. Similarly, the locations of the opposing team and the ball are recorded in the second and third channels. In layers four to nine the velocities of all players and of the ball are recorded by entering the speed in both longitudinal and transverse directions at the respective spatial locations.

The resulting representation is then processed by a CNN with three layers. The output of the CNN is fed to a fully-connected layer, which estimates the value of each position (as an additional input, the fully-connected layer receives the information of ball possession). Comparing these estimated values to actual game results, the authors conclude the usefulness of the learned evaluation for soccer analysis.

Among other things, the spatial representation helps to counter a permutation problem: If the data was processed directly with a fully-connected neural network, the order of players in the data would affect the evaluation of a position. Representing the information of players of a team in only one channel counteracts this problem. As an alternative to the approach described here, representing positional data as graphs and processing positional data with *permutation equivariant* models, such as *graph neural networks* (e.g., Yeh et al., 2019) or *transformers* (cf. Rudolph & Brefeld, 2022), appears promising.



■ Fig. 22.3 The data representation of positional soccer data suitable for processing with a CNN as proposed by Dick and Brefeld (2019)

References

- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition* (pp. 248–255).
- Dick, U., & Brefeld, U. (2019). Learning to rate player positioning in soccer. *Big Data*, 7, 71–82.
- Fassmeyer, D., Anzer, G., Bauer, P., & Brefeld, U. (2021). Toward automatically labeling situations in soccer. *Frontiers in Sports and Active Living*, 3, 725431.
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36, 193–220.
- Gerke, S., Müller, K., & Schäfer, R. (2015). Soccer Jersey Number Recognition Using Convolutional Neural Networks. *IEEE International Conference on Computer Vision Workshop*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Johnson, S., & Everingham, M. (2011). Clustered pose and nonlinear appearance models for human pose estimation. In *IEEE conference on computer vision and pattern recognition*.
- Kayhan, O. S., & van Gemert, J. C. (2020). On translation invariance in CNNs: Convolutional layers can exploit absolute spatial location. In *IEEE conference on computer vision and pattern recognition*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 2278–2324.
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In *International conference on machine learning*.
- Rudolph, Y., & Brefeld, U. (2022). Modeling conditional dependencies in multiagent trajectories. In *International conference on artificial intelligence and statistics*.
- Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *IEEE conference on computer vision and pattern recognition*.
- Yeh, R. A., Schwing, A. G., Huang, J., & Murphy, K. (2019). Diverse generation for multi-agent sports games. In *IEEE conference on computer vision and pattern recognition*.



Transfer Learning

Henrik Biermann

Contents

23.1 Example Sport – 194

23.2 Background – 195

23.3 Applications – 196

References – 199

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>.
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>.

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Traditional machine learning (e.g., using Deep Neural Networks, as discussed in ► Chap. 22) generally requires large amounts of data (Big Data) and powerful computing systems (computational clusters).
- Transfer learning allows the utilization of pre-trained models (e.g., from image processing) for solving new problems.
- This may require a transformation of data representation (domain) if necessary.
- Transfer learning can lead to new insights into problem-solving.

23.1 Example Sport

In addition to automated machine learning, the concept of transfer between different tasks (or skills) in sports is of special significance. For instance, the human ability to translate strength training performance into good performance in various sports has been studied. In computer science, this principle is used to apply machine-learned models to new tasks. To illustrate this, an example from football tactical analysis is presented. The basis for transfer learning often involves powerful models from image processing that can reliably recognize various objects in images. We aim to use one of these models through transfer learning for analyzing (rating) ball possession phases. To fully exploit the potential of the base model, it requires a graphical representation of the data (an image or video) as input. Various representations of ball possession phases can be used, including a video frame (e.g., television footage) or converting position data (see ► Chap. 6) into a graphical representation. After these transformations, the rating of ball possession phases can be performed. Comparing the results of different representations and classical (feature-based) methods can be particularly interesting. Poor results in video frames may indicate that the large redundancy present in the videos hampers the models, whereas good results in abstract two-dimensional images suggest that most

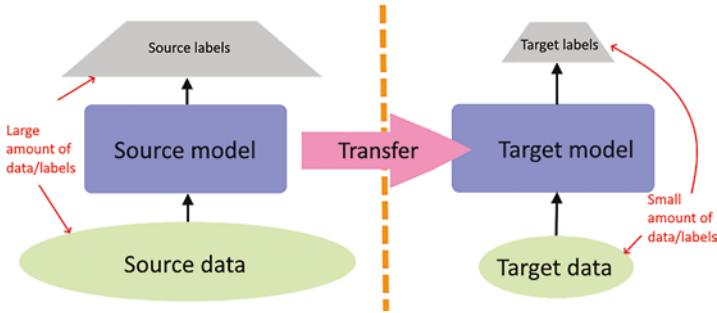
relevant information from the position data is present in this representation. Lastly, comparing with feature-based methods can demonstrate potential advantages of transfer learning.

23.2 Background

In recent years, various fields have seen an increasing adoption of machine learning methods, where models with a large number of parameters (several millions) are tailored to specific tasks. This process involves training the models for the given task, which requires a substantial amount of data (Big Data) to ensure their applicability to unseen data, as well as powerful computational resources (rechencluster) to conduct the training within a practical timeframe. These conditions are usually only met by large companies. For smaller groups (research teams) that often lack access to a significant amount of task-specific data and powerful computing resources, training such large models becomes impractical. However, while the data and computational capabilities of large companies are not made publicly available, there is an opportunity to download pre-trained models from the internet. An example of such a model is ImageNet (Russakovsky et al., 2015), a deep neural network (see ► Chap. 10), which can reliably recognize over 1000 object categories in more than one million images.

This is where transfer learning comes into play. Based on this principle, a model that has already been designed (and trained) for a specific task can be used for a new task. The layered structure of the models, consisting of different independent layers, can be exploited. In the case of ImageNet, the first layer of the network receives an image (e.g., an orange), processes it, and then passes the result to the next layer of the network. This process continues until the final layer outputs an object class. Consequently, in the final layer of the network, the object class is determined based on the output of the penultimate layer. The output of the penultimate layer can be considered as a (high-dimensional) representation (feature vector) of the original input image (orange), containing all necessary information. Since an individual layer (layer) of a deep neural network can be roughly compared to a regression (Dreiseitl & Ohno-Machado, 2002), where data points in a (high-dimensional) space are divided into different subgroups, it can be assumed that the features created by the initial layers are domain-independent. For the example of ImageNet, this means that the first layers perform image processing operations comparable to “classical” image processing (edge detection, convolution, etc.). Therefore, the pre-trained neural network can not only reliably recognize the 1000 object categories but also (domain-independently) extract relevant information from an image.

As a result, in addition to the cost-effective use of very large models, transfer learning also offers the opportunity to gain profound insights into the structure of a problem through domain transfer. The transformation of data representation plays a crucial role in this process. This is particularly interesting for complex contexts, such as those encountered in sports game analysis (■ Fig. 23.1).



■ Fig. 23.1 Overview of the idea of transfer learning

Definition

Transfer learning describes the process of repurposing a machine-trained model for a new task. This enables the utilization of large and powerful models for specific tasks, even in the absence of powerful computational systems and large datasets for the given task.

23.3 Applications

► Example 1

As previously demonstrated, deep neural networks from image processing are commonly used as the base model. Thus, there are examples where transfer learning is used to automatically recognize the content of visual data. In a study by Russo et al. (2019), a pre-trained network from the ImageNet dataset was employed for the automatic recognition of sports videos. The authors compiled a video dataset comprising television images from a total of 15 sports (football, rugby, table tennis, volleyball, basketball, cricket, etc.). Using this dataset, they trained a model to automatically assign the correct sport to the videos. The study's results show that transfer learning from the pre-trained model improves the previous gold standard accuracy of 96% to a perfect accuracy of 100%, showcasing the powerful image-processing elements of the pre-trained model and the advantages of transfer learning. ◀

► Example 2

Another application of transfer learning in video images comes from a study by De Campos et al. (2013). The authors present a model capable of automatically recognizing events in videos. The model's hierarchical structure allows abstraction at different semantic levels. The authors demonstrate this with an application to television images of tennis matches. In the lowest (low-level) layer, individual shots are initially recognized. This is followed by field markings, player detection, and ball recognition in upper lay-

ers. Finally, the top hierarchical layer detects events, enabling automatic annotation of events such as serves, shots, or ball hits in tennis, and even automatic recognition of the current score. This particular model structure can also be applied for transfer learning. The authors define anomalies that can be detected at each hierarchical level. An anomaly in one of the lower hierarchical levels could indicate that the shown video is not from a tennis match, while an anomaly in an upper level could suggest that too many players are detected on the field. If an anomaly (different sport) is detected at the lowest level, linear transformations are automatically triggered to adapt the model to the new environment. By doing this, the authors successfully transfer the model to the sport of badminton. ◀

► Example 3

Various transfer learning methods deal with tactical analysis in sports. As a new method of graphical representation of positional data in football, Visual Rhythms (Rodrigues et al., 2017) were introduced. This graph-based approach creates graphs from positional data and extracts specific features from them. As these features vary over time (see time series, ► Chap. 26), they can be transformed into visual representations. These Visual Rhythms can be imagined as sequences of pixels, where the image's length encodes the duration of a situation. The feature's value can be represented by changing colors (similar to a heatmap). Furthermore, different Visual Rhythms for different features can be stacked to create a two-dimensional image reflecting the evolution of various features over time. This powerful and concise analysis tool provides a visual representation of positional data, enabling detailed evaluation of various tactical concepts in football. However, the specific application of Visual Rhythms for machine-based football analysis has not been thoroughly investigated. ◀

► Example 4

Another example involves an adaptation of the well-known AlphaGo algorithm by Google DeepMind (Silver et al., 2016). Its victory against professional Go player Lee Sedol was perceived as highly significant by the community at the time. However, just 2 years later, a variation called the AlphaZero algorithm (Zhang & Yu, 2020) was introduced. This algorithm, based on the original AlphaGo, can now play chess and shogi (a Japanese chess variant) in addition to Go. Another unique feature of AlphaGo is that it is trained solely by playing against itself and does not require external data (aside from the rules of the games). In a showcase match against the then-gold-standard chess engines Stockfish 8, AlphaGo convincingly won 64 out of 100 games, demonstrating the adaptability of the original algorithm and the similarity between Go and chess. ◀

► Example 5

Transfer learning is also increasingly applied in other domains, such as in medicine, where networks from the ImageNet Challenge (Russakovsky et al., 2015) can be transferred to the medical domain. An example is the “Medical Segmentation Decathlon”

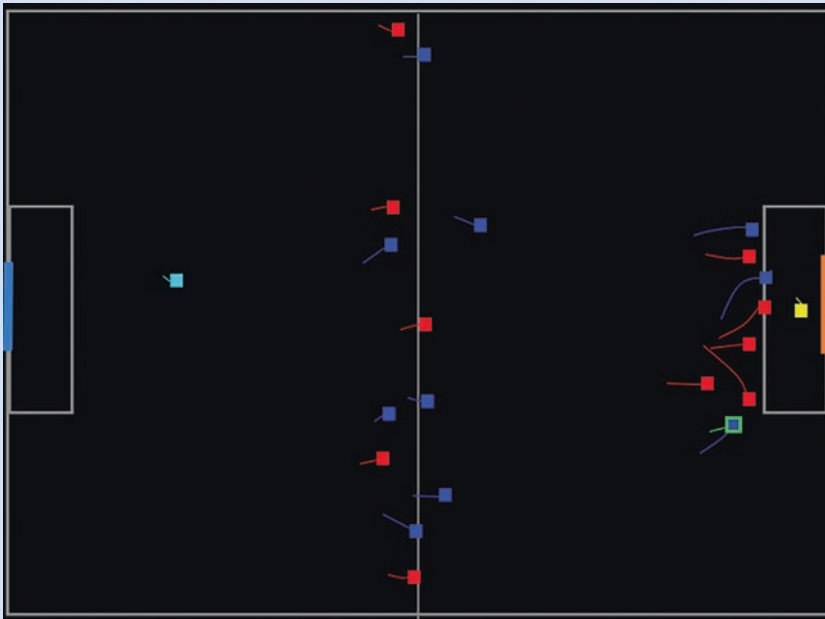
challenge (Antonelli et al., 2021), in which a medical image dataset with corresponding labels was released. By applying transfer learning in this challenge, medical disease detection and prevention are significantly improved. For example, the model can decide for an image whether a healthy vessel or a harmful tumor is depicted. Challenges in this problem include the relatively low amount of data and the relative similarity between the target classes. ◀

? Questions for the Students

1. What are the advantages of transfer learning compared to a “classical” machine learning approach?
2. Why is the last layer of the Neural Network removed during transfer learning?

Study Box

In a study on the application of transfer learning in football analysis, Wagenaar et al. (2017) asked how attacking sequences in football can be reliably evaluated. They observed attacking sequences that either resulted in a goal-scoring opportunity or a loss of possession. A GoogLeNet model with over four million parameters was used as the base model for transfer learning, trained either on the ImageNet dataset (Russakovsky et al., 2015) or solely on the available dataset. To analyze positions with this network, a visual representation was defined. For each time point, three-channel (two-dimensional) RGB images with a size of 256×256 pixels were generated from the positional data. The field was displayed as a black square with some important field markings (sidelines, halfway line, and both penalty areas). Players and the ball were represented as “blobs” (colored pixels) based on their coordinates in the positional data. Both goals were also represented as colored lines, with the ball shown in green, the home team in dark blue (players), cyan (goalkeeper), and blue (goal), and the away team in red (players), yellow (goalkeeper), and orange (goal). To depict the movements during attacking sequences, the trajectories of players within the last 2 s were shown as lines adjacent to the blobs. For the specific rating task of attacking sequences, the authors compared the best results of GoogLeNet with those of a classical K-nearest-neighbor (KNN) approach. Transfer learning showed an improved accuracy of 10% (67.1%) compared to the KNN approach (57.3%) (■ Fig. 23.2).



■ **Fig. 23.2** Visual representation of positional data following the example of Wagenaar et al. (2017)

References

- Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B. A., Litjens, G., ..., & Cardoso, M. J. (2021). The medical segmentation decathlon. arXiv preprint arXiv:2106.05735.
- De Campos, T. E., Khan, A., Yan, F., FarajiDavar, N., Windridge, D., Kittler, J., & Christmas, W. (2013). A framework for automatic sports video annotation with anomaly detection and transfer learning. Machine learning and cognitive science, collocated with EUCOGIII.
- Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5–6), 352–359.
- Rodrigues, D. C. U. M., Moura, F. A., Cunha, S. A., & Torres, R. D. S. (2017, February). Visualizing temporal graphs using visual rhythms—a case study in soccer match analysis. In *International conference on information visualization theory and applications* (Vol. 4, pp. 96–107). SciTePress.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 211–252.

- Russo, M. A., Kurnianggoro, L., & Jo, K. H. (2019, February). Classification of sports videos with combination of deep learning models and transfer learning. In *2019 international conference on electrical, computer and communication engineering (ECCE)* (pp. 1–5). IEEE.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–489.
- Wagenaar, M., Okafor, E., Frencken, W., & Wiering, M. A. (2017, February). Using deep convolutional neural networks to predict goal-scoring opportunities in soccer. In *International conference on pattern recognition applications and methods* (Vol. 2, pp. 448–455). SCiTePress.
- Zhang, H., & Yu, T. (2020). AlphaZero. In *Deep reinforcement learning* (pp. 391–415). Springer.



Random Forest

Justus Schlenger

Contents

24.1 Example Sport – 202

24.2 Background – 203

24.3 Applications – 204

References – 207

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>.
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>.

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com.

Key Messages

- Random forest is a machine learning algorithm based on the use of a large number of decision trees for predictions
- The method can be used without much fine tuning and settings (out of the box)
- The Random Forest combines the simplicity and intuitiveness of decision trees with the complexity and flexibility of ensemble methods
- Especially in the study of performance indicators in sports the binary decision character lends itself

24.1 Example Sport

The Random Forest (RF) algorithm, developed by Breiman (2001), is a machine learning method that enjoys great popularity in data science. This can be justified, among other things, by its fundamental building block, the decision tree. The structure of a decision tree mirrors the structure of decision processes, such as those that play an important role in sports. This could be the decision of a soccer coach about a player change, a player recruitment (Koenigstorfer & Wemmer, 2019) or the tactical decision of a tennis player regarding the chosen stroke. In the field of computer and data science, the RF algorithm is counted among the machine learning or artificial intelligence methods. This is because RF, like other methods in this category, will attempt to generate a model using known data, which will then allow predictions to be made using unknown or future data. The sports computer scientist can make use of these predictions, for example, in relation to the study of sports betting. For the exemplary examination of a single basketball game, data of both teams regarding certain parameters have to be collected. These could be possible performance indicators, such as the body sizes, the running distances in past games or the market values of the players. The multitude of decision trees is now trained (built) with the corresponding parameters based on results of past games. Subsequently, the model can be applied to future game data, allowing the decision trees to cast individual votes for the final outcome of the game. Using

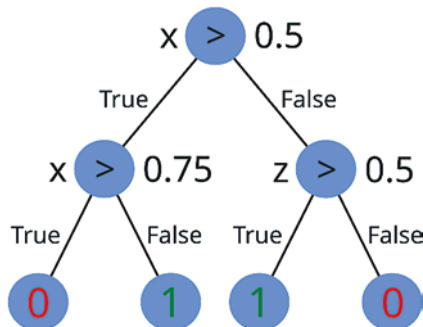
a majority vote, the sports computer scientist can now not only make predictions for the basketball game, but also investigate the importance of the individual parameters (possibly KPIs) used in the model.

24.2 Background

The operation of the RF algorithm is best illustrated by a single binary decision tree. By convention, one usually starts at the top at the “root” and works through the branches to one of the “leaves” at the bottom of the tree. The branches are called “nodes” and the lines in between are called “branches”. ■ Figure 24.1 shows a highly simplified version of a decision tree. At the root and at the nodes certain decision rules are placed, which contain either discrete categories or thresholds. If the model of the decision tree has already been trained, it can be used to predict a class or a numerical value for units under consideration. The former is a classification tree and the latter is a so-called regression tree (Myles et al., 2004). Therefore, decision trees, justified by Breiman et al. (1984), are also called “CARTs” (Classification and Regression Trees).

In the example in ■ Fig. 24.1, the binary groups 0 and 1 are assigned on the basis of decision rules regarding the parameters x and z . Using the names from statistics, the classes 0 and 1 represent the dependent variable and the parameters x and z the independent variables. Thus, in this case it is a classification tree. However, as mentioned earlier, a decision tree must first be built to be useful for classification or regression tasks. To do this, labeled training data must be available, that is, those data from past events in which the final results are already known. These types of machine learning algorithms belong to “supervised learning” methods. In contrast, methods of unsupervised learning are characterized by missing labels.

The algorithm for creating a decision tree runs in such a way that the entire training data set (starting at the root) is split at each node in such a way that groups are created that are as homogeneous as possible. This process can be repeated as often as needed until the dataset has been split optimally with respect to the given labels or the maximum number of branches has been reached. This maximum number of branches can be determined by the data scientist and is called the “depth” of the decision tree.



■ Fig. 24.1 Sketch of simple classification tree

However, as it turns out, a single decision tree has a very low flexibility towards unknown data and tends to incorporate the random variance of the data for decision making (Biau & Scornet, 2016). This phenomenon is called “overfitting” and results in a model that is not able to detect actual relevant patterns in the data. For this reason, the Random Forest makes use of an arbitrarily large ensemble of decision trees, all of which have different and random subsets of the data and parameters at their disposal. When such a training process is completed, the resulting forest can be used just like a single tree to make predictions about unknown data. Taking a classification problem as an example, the prediction of the Random Forest algorithm corresponds to the class that received the most “votes” of the individual trees. Thus, the Random Forest algorithm combines the intuitive and simple structure of decision trees with the flexible and robust nature of ensemble methods, which are used in many state of the art machine learning methods (Hastie et al., 2009). Last but not least, after using the algorithm, those parameters can be filtered out that helped to achieve the comparatively best prediction overall.

Definition

The Random Forest is an ensemble machine learning method dedicated to classification or regression tasks by constructing numerous slightly different and uncorrelated decision trees.

24.3 Applications

► Example 1 Decision Tree -> RF

To outline the use of a single decision tree, the example of the already mentioned basketball game is used. The decision tree shown in [Fig. 24.2](#) is supposed to predict the victory or defeat of any team in the upcoming game based on the parameters of ball possession (BP) and running distance (LD) from previous games. In doing so, the cor-

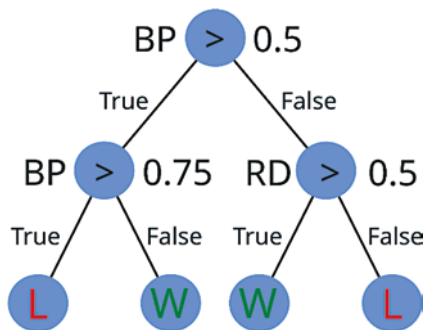


Fig. 24.2 Exemplary representation of a decision tree using the example of score prediction in basketball. The parameters ball possession (BP) and running distance (LD) are used as decision support for the prediction of victory (S) and defeat (N)

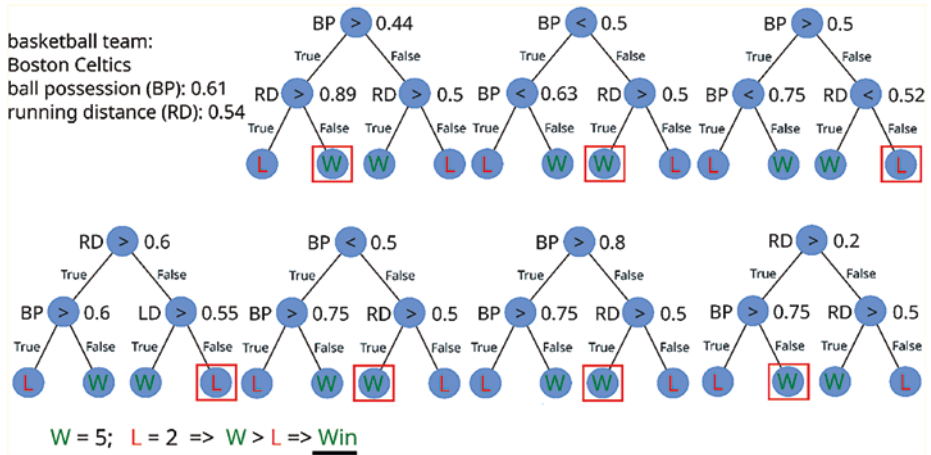


Fig. 24.3 Random forest with 7 decision trees using the basketball game as an example. The trained RF is used to predict the team “Boston Celtics”, which shows an average possession of 61% and an average percentage of total running distance per game of 54%

responding data from past games are collected and injected into the model. This results in the decision tree shown in [Fig. 24.2](#).

When a team is considered, the classification starts at the root, following either the left branch if the average possession is above 50% or the right branch if the average possession is below 50%. Assuming our considered team has an average possession of 60%, another node follows which has a threshold at 75%. If our team is below this value (False) a win is predicted. If this classification now takes place with an already trained random forest instead, each examined unit is promoted through all decision trees in parallel. In [Fig. 24.3](#), hypothetical values have been chosen for the Boston Celtics team. This results in a prediction of win (W) or loss (L) for the upcoming game in each of the different decision trees.

Here we now check which of the two classes was assigned more frequently, in order to finally use the more frequent class as a prediction by majority rule. ◀

► Example 2 RF

In this example, we consider a study by Smithies et al. (2021), which aims to find performance indicators of the e-sport “Rocket League” that predict expertise and success of e-sports players. In the video game “Rocket League,” each player controls a rocket-powered sports car and attempts to launch a large ball into the opponent’s goal, similar to the partially familiar “autoball.” In the process, the sports car can also drive along walls and fly through the air by means of the rocket propulsion. The influence of metrics such as “average speed”, “number of shots on goal” and “time spent in the air” on the match result (success) and player rank (expertise) is tested. Here, the analysis of 20,000 matches revealed that “shots on goal taken”, “preventing opponent’s shots on goal” and “ball saves” best predicted the final outcome of the match. On the other hand, for the prediction of player rank (league system: bronze, silver, gold...), the metrics “time spent on the ground” and “time in high speed”, for example, provided the highest accuracy. ◀

► Example 3 RF + Poisson Regression

In the study by Groll et al. (2019), a hybrid version of Random Forest was used together with Poisson regression to predict outcomes of the 2018 FIFA World Cup. Poisson models were used to create a team ranking that incorporates recent team performance more than longer past performance. Thus, the current team strength should be determined as accurately as possible. This ranking could be used as further covariates for the Random Forest Model. The model was trained with data from the FIFA World Cups 2002–2014 and the predictive power was higher than other methods known at that time including betting odds. This also shows that the RF algorithm can be easily combined with other methods. ◀

► Example 4 RF

In many sports, competitive sport is a business in the multiple millions. One element that is often overlooked when recruiting players is the risk of injury. Serious injuries can not only waste huge amounts of money from the club side, but also destroy entire careers and thus the future of young athletes. In a study by Jauhiainen et al. (2021), predictors for injuries are identified using the Random Forest algorithm, so that both athletes and clubs can obtain a better assessment of injury risk. Highly relevant parameters are gender, mobility of the biceps femoris, body mass index (BMI) and height. Thus, clubs or even the players themselves can take appropriate measures regarding these predictors. Even with unchangeable predictors such as height or gender, athletes can be monitored and coached in a more targeted manner. ◀

24

? Questions for the Students

1. Why do not all decision trees usually cast the same vote in the Random Forest algorithm (classification)?
2. Give an example from the field of sports where RF could be used. Give five different possible parameters for this scenario.

Study Box

In a study on soccer by Jamil et al. (2021), a set of key performance indicators (KPIs) was used to try to distinguish Champions League goalkeepers from non-Champions League goalkeepers. In this study, on a dataset with an observation number of $n = 14,671$ match scenarios, the Random Forest algorithm was used for classification, among other methods. As a result of this binary classification, it is shown that the main difference is not so much in number of shots held as in short passing with the foot. The Random Forest algorithm gave an accuracy of 0.66 which means that 66% of all predictions were correct. This is significantly above the 50% threshold (baserate) that would be achieved by simple guessing. Predicting known player levels to identify key performance indicators is one of the most common applications of Random Forest in sports science. This framework can be adapted to different questions without any problems. Only the (input) parameters (independent variables) and the coding of the target variable (dependent variable) have to be modified.

References

- Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), 197–227.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks.
- Groll, A., Ley, C., Schaubberger, G. & Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, 15(4), 271–287. <https://doi.org/10.1515/jqas-2018-0060>.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random forests. In *The elements of statistical learning* (pp. 587–604). Springer.
- Jamil, M., Phatak, A., Mehta, S., Beato, M., Memmert, D., & Connor, M. (2021). Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men's football. *Scientific Reports*, 11(1), 1–7.
- Jauhiainen, S., Kauppi, J. P., Leppänen, M., Pasanen, K., Parkkari, J., Vasankari, T., et al. (2021). New machine learning approach for detection of injury risk factors in young team sport athletes. *International Journal of Sports Medicine*, 42(2), 175–182.
- Koenigstorfer, J., & Wemmer, F. (2019). What makes sports clubs successful at recruiting and retaining members from the perspective of managers? Results from a random forest analysis. *Journal of Global Sport Management*, 7, 644–663.
- Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics*, 18(6), 275–285.
- Smithies, T. D., Campbell, M. J., Ramsbottom, N., & Toth, A. J. (2021). A Random Forest approach to identify metrics that best predict match outcome and player ranking in the esports Rocket League. *Scientific Reports*, 11(1), 19285.



Statistical Learning for the Modeling of Soccer Matches

Gunther Schauberger and Andreas Groll

Contents

- 25.1 Example Sport – 210
- 25.2 Background – 211
- 25.3 Applications – 212
- References – 214

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>.
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>.

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- For the modeling of soccer matches, the response/target variable can be defined in different ways. Depending on the choice of the response variable, different approaches of statistical learning or machine learning are suitable for the modeling and prediction of soccer matches
- While at the beginning mainly classical regression methods were used, in recent years machine learning methods such as extreme gradient boosting or random forests have been applied more frequently
- Machine learning methods could in particular improve the prediction quality for new matches in the future, but are also more complex and more difficult to interpret

25.1 Example Sport

In sports, and especially in soccer, by now a large amount of data is collected and analyzed. An important goal of this evaluation is the modeling or prediction of individual matches with the help of so-called *statistical learning*. One is often interested in the match outcome, which can be captured mathematically in different ways, and how it depends on different covariates (features). For example, the match outcome can be considered as the precise result in goals or as an ordinal variable with the three categories "win team A", "draw" and "win team B". Typically, the covariates are characteristics of the two competing teams, such as their respective market values, standings, or previous performances. Furthermore, the general conditions of the respective match (e.g. weather, home advantage, point vs. friendly match, ...) can also be included in the modeling. Principally, the modeling approaches described here can also be applied to other sports, although there are always special modeling requirements depending on the sport. For example, the outcomes of tennis matches can simply be considered as binary variables (win player A or win player B). The modeling of basketball matches, on the other hand,

differs from the modeling of soccer matches by the much higher numbers of scores or points.

25.2 Background

The situation described above, where a certain response variable (in our case the outcome of the match) is to be modeled with the help of covariates, represents a classic case of statistical learning, or more precisely so-called *supervised statistical learning*. The term “supervised” means here that for the individual observations (usually the matches) not only covariates are available, but also the corresponding response. In contrast, in the case of unsupervised learning, the response is unknown or not available. Typical approaches from supervised learning are (linear) regression and classification, a typical example of unsupervised learning are clustering methods. A comprehensive introduction to statistical learning can be found in James et al. (2021).

In most cases, the basic question for the choice of the response variable is whether the number of goals or the ordinal match outcome should be chosen. The more common choice is to model the number of goals. Since we are dealing with discrete count data (0, 1, 2, 3, ...), modeling using the so-called *Poisson distribution* is a good choice instead of the commonly used normal distribution. For simplicity, the two numbers of goals belonging to one match (i.e., the goals scored by Team A and Team B) are often assumed to be (stochastically) independent (Groll et al., 2015). However, this independence should only be understood as conditional independence (given the information in the different covariates). The consequence of this assumption is that the two goal counts of a match can be treated as two (conditionally) independent observations in the data set. For example, a dataset of 100 matches would therefore contain 200 observations. The alternative is to directly use the tuple of the two numbers of goals of a match as the response, which requires bivariate modeling or distributional assumptions (Karlis & Ntzoufras, 2003; Groll et al., 2018). For example, this can be solved using so-called copula regression (van der Wurp et al., 2020).

An ordinal match result would usually only contain the information “win team A”, “draw” and “win team B” or “defeat team A”, respectively, which can be encoded with the values 1, 2 and 3. Where necessary, the target variable may also reflect the margin of the victory (or defeat), resulting in more than three categories. In this context, a match represents a single observation. The response variable then is an ordinal variable with the possible values 1, 2 or 3. In a regression context, so-called ordinal regression is used for modeling (Schauberger et al., 2018).

In the simplest case, all of the possibilities mentioned above can be modeled by a linear regression model, where the response is modeled as the sum of the linear effects of the individual covariates. Usually, one unknown regression parameter (effect) is estimated per covariate (see ► Chap. 16). Sometimes a large number of (potential) covariates can be involved in the modeling so that a very large number of parameters would have to be estimated, with the consequence that ordinary methods would be unstable or even unfeasible. In this situation, so-called regularization methods, in particular variable selection techniques such as Lasso (Tibshirani, 1996) or boosting (Friedman, 2001), can be helpful.

In this chapter, we will focus exclusively on the modeling of individual matches. However, this can in a second step serve as the basis for the modeling or simulation of entire tournaments (e.g. World Cups and European Championships) or even national championships. For a tournament, the entire course of the tournament is simulated very often (e.g. one million times). The results of the individual matches are based on a fitted model trained on a (learning) data set from previous tournaments and are randomly drawn by the help of the corresponding estimates. Initially, this results in, for example, one million individual world champions. Based on these, for all teams, the corresponding relative frequencies can be determined and winning probabilities be derived. The repeated simulation of the tournament automatically accounts for any special features of the groups or the tournament schedule (Groll et al., 2015).

25

Definition

Statistical learning (Hastie et al., 2009) covers a large number of methods designed to extract information from data. In so-called supervised learning, there is a response variable that is to be explained and predicted using various covariates. The relationship between the target variable and the covariates is observed on a learning data set and, if necessary, is transferred to new observations to predict their responses.

25.3 Applications

In the following, we provide three examples of the three aforementioned ways of defining the response in soccer matches and modeling it accordingly. Example 1 presents applications where the individual numbers of goals were modeled (conditionally) independently, while Example 2 describes applications of bivariate modeling. Finally, Example 3 focuses on models for ordinal response variables.

► Example 1

For the prediction of the 2014 FIFA World Cup (WC), Groll et al. (2015) considered the modeling of matches from previous FIFA WCs, based on a learning dataset covering all matches of the WCs 2002–2010. Following the model of Dixon and Coles (1997), a Poisson model was estimated in which the numbers of goals of a match were assumed to be conditionally independent. However, their original (rather simple) model was extended to include team-specific covariates. To handle the large number of parameters involved, a Lasso regularization was used to estimate the model. Variations of this model were used to model and predict the 2019 IHF Handball World Cup (Groll et al., 2020). For the prediction of the 2018 World Cup, the Poisson model was replaced by a so-called *random forest* (a special machine learning model, see Groll et al., 2019; see also Study Box). Further extensions of this model were then used in the prediction of the 2019 FIFA World Cup in women's soccer (Groll et al., 2019) and the 2020 UEFA European Championship (EURO) in men's soccer (Groll et al., 2021). In the latter work,

the prediction performance of the random forest approach was compared to so-called extreme gradient boosting (xgboost, see Chen and Guestrin, 2016). ◀

► Example 2

In the context of modeling and predicting the UEFA EURO 2016, Groll et al. (2018) investigated the validity of the assumption of conditional independence of the two numbers of goals of a match. For this purpose, they used a regression model based on the bivariate Poisson distribution, which was first introduced by Karlis and Ntzoufras (2003) for modeling soccer matches. Estimation of this model was performed via boosting. The bivariate modeling has not shown any advantages over the simpler (independence-based) modeling here.

In addition to the specific modeling assumption of a bivariate Poisson distribution, also copula models can be used, where the marginal distributions are based on univariate Poisson distributions. Such approaches were applied in the modeling of FIFA World Cup matches by van der Wurp et al. (2020) and Van der Wurp and Groll (2021). ◀

► Example 3

In Schauburger et al. (2018), the influence of so-called match-specific variables on the outcome of matches in the German soccer Bundesliga was investigated. Examples of such match-specific covariates are the ball possession percentage or the running distance of the two teams. For this purpose, an ordinal variable with five categories was used as the response variable in order to be able to differentiate between higher and lower victory margins in addition to draws.

In Schauburger and Groll (2018), a special ordinal variant of the random forest was used to model FIFA World Cup matches, where the match outcome was considered as an ordinal variable with the categories “win team A,” “draw,” and “win team B” and contrasted with different models for univariate goal modeling. ◀

? Questions for the Students

1. What are the different ways of mathematically encoding or modeling the results of soccer matches?
2. Name possible statistical learning or machine learning approaches that have already been used to model soccer matches!

Study Box

For the prediction of the FIFA World Cup 2018, Groll et al. (2019) created a training dataset including all matches of the previous World Cups 2002–2014. Here, the individual numbers of goals were considered as (conditionally) independent observations of the response variable, i.e., resulting in two observations per match (cf. situation of Example 1). Various variables were collected for the competing teams, namely economic factors (such as the GDP of the respective country), sportive factors (such as the respective position in the FIFA world rankings), variables repre-

senting the home advantage (e.g., a dummy variable for whether one of the two teams in a match is the home team, i.e., the host country of the World Cup), variables related to the team structure (such as the average age of the respective team), and variables related to the respective coach (e.g., the length of his tenure to date). Furthermore, another sportive variable was added, which has a special role, since it was not directly available, but had to be estimated by a separate statistical model. The variable represents ability parameters which, for each team, reflect their current strength (at the time of the start of the respective World Cup) and were estimated based on historical matches (all matches from the last 6 years with a time weighting factor).

A *hybrid* random forest model was then fitted on this training data set. The model is referred to as hybrid because team-specific ability parameters, which themselves were derived from another statistical model, were added to the set of covariates. Based on the estimated model fit and the teams' covariates for the upcoming 2018 World Cup, the entire course of the tournament was then simulated 100,000 times. Thus, winning probabilities for all 32 participating teams could be determined. In retrospect, the prediction performance of this hybrid random forest model was compared on all 64 matches of the 2018 World Cup to various other statistical modeling approaches, in particular also an ordinary random forest (i.e., without the hybrid team-specific abilities) as well as the bookmakers' betting odds. The hybrid random forest model performed best and achieved very satisfying results overall.

References

- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794).
- Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46, 265–280.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29, 337–407.
- Groll, A., Heiner, J., Schaubberger, G., & Uhrmeister, J. (2020). Prediction of the 2019 IHF world men's handball championship—A sparse Gaussian approximation model. *Journal of Sports Analytics*, 6(3), 187–197.
- Groll, A., Hvattum, L. M., Ley, C., Popp, F., Schaubberger, G., Van Eetvelde, H., & Zeileis, A. (2021). Hybrid machine learning forecasts for the UEFA EURO 2020. arXiv preprint arXiv:2106.05799.
- Groll, A., Kneib, T., Mayr, A., & Schaubberger, G. (2018). On the dependency of soccer scores—A sparse bivariate Poisson model for the UEFA European football championship 2016. *Journal of Quantitative Analysis in Sports*, 14(2), 65–79.
- Groll, A., Ley, C., Schaubberger, G., & Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, 15(4), 271–287.
- Groll, A., Ley, C., Schaubberger, G., Van Eetvelde, H., & Zeileis, A. (2019). Hybrid machine learning forecasts for the FIFA women's world cup 2019. arXiv preprint arXiv:1906.01131.

- Groll, A., Schauburger, G., & Tutz, G. (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA world cup 2014. *Journal of Quantitative Analysis in Sports*, 11, 115–197.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. Springer.
- Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 5, 381–393.
- Schauburger, G., & Groll, A. (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling*, 18(5–6), 460–482.
- Schauburger, G., Groll, A., & Tutz, G. (2018). Analysis of the importance of on-field covariates in the German Bundesliga. *Journal of Applied Statistics*, 45(9), 1561–1578.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288.
- Van der Wurp, H., & Groll, A. (2021). Introducing LASSO-type penalisation to generalised joint regression modelling for count data. *AStA Advances in Statistical Analysis*, 107, 127–151.
- van der Wurp, H., Groll, A., Kneib, T., Marra, G., & Radice, R. (2020). Generalized joint regression for count data: A penalty extension for competitive settings. *Statistics and Computing*, 30, 1419–1432.



Open-Set Recognition

Ricardo da Silva Torres

Contents

26.1 Example Sport – 218

26.2 Background – 218

26.3 Applications – 220

References – 221

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>.
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>.

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

26

Key Messages

- In several practical pattern recognition problems, machine learning solutions should be able to inform if a sample does not belong to any of the classes considered for training
- In Sports Science, several practical problems should be modelled as open-set recognition tasks
- The use of open-set recognition solutions in Sports Science applications is a problem overlooked in the literature

26.1 Example Sport

Computer Vision is one of the areas of Computer Science dedicated to the interpretation and understanding of multimedia data. Several Computer Vision approaches rely on the use of machine learning methods. Naik et al. (2022) overview recent literature in the area focusing on Sports Science. In several of the surveyed applications, open-set recognition (OSR) algorithms could be investigated. Two examples will be used to illustrate the use of an OSR formulation. The first one refers to the automatic classification of tactics employed by soccer teams, while the second refers to the problem of recognizing relevant players' actions in soccer videos.

26.2 Background

Huge collections of sport-related data have been created due to technological innovations related to the development of monitoring systems and the availability of low-cost powerful storage and processing computer systems. The adequate analysis of the available data has been recognized as a valuable asset in supporting sports analysis aiming at better-informed decision-making (Goes et al., 2021; Rein & Memmert, 2016).

One relevant trend in the area of Sports Science refers to the use of data-driven methods for supporting knowledge discovery. Among the most used approaches, machine learning methods have been successfully used. Machine learning (ML) is a subset of Artificial Intelligence technologies used to learn from data, aiming to support prediction and inference tasks or the identification of associations among data (Enholm et al., 2021). Existing ML methods are often grouped into four different categories: supervised, unsupervised, semi-supervised, and reinforcement learning approaches. Supervised approaches (e.g., classifiers) assume the existence of labels (class or category) associated with samples used in the training process. In this case, models trained using a labeled collection are expected to generalize to unseen samples at the test phase, i.e., a classifier is considered effective or accurate if its use leads to correct predictions of the labels associated with test samples.

What if the testing sample does not belong to any of the categories considered in the training phase? In this case, the ML method should not assign labels considered in the training set; it should somehow inform that the testing sample belongs to an unknown class. In fact, in several practical problems, the number of classes or categories to be considered in the design and implementation of ML methods can not (or should not) be defined in advance.

Definition

In machine learning, most of the time, we do not need, do not have access to, or are not aware of all possible classes to consider at training time (de Oliveira Werneck et al., 2019; Mendes Júnior et al., 2017; Neira et al., 2018). For instance, when classifying whether or not a video contains a particular action of an athlete or a referee, we might have training examples of only positive cases, i.e., videos associated with a predefined set of possible actions (e.g., Naik et al., 2022). Open-set recognition (OSR) refers to the problem of identifying the unknown classes during testing while maintaining performance on the known classes (Oza & Patel, 2019).

The first initiatives towards the definition of OSR algorithms relied on the extension of consolidated classification methods e.g., Support Vector Machines as explored by Scheirer et al. (2012), and class proximity information as investigated by Mendes Júnior et al. (2017), and Cardoso et al. (2017). More recently, deep learning approaches became a trend. Earlier initiatives focused on predicting unknown samples in the final layer of proposed architectures (Bendale & Boult, 2016; Ge et al., 2017; Liang et al., 2017. Liang et al. (2017), for example, presented ODIN, an approach that explores temperature scaling and small perturbations in the detection of out-of-distribution samples. Bendale and Boult (2016) introduced a reweighting scheme to redefine the output probabilities to detect unknown samples. The proposed formulation, known as OpenMax, was extended by Ge et al. (2017), who incorporated training procedures involving synthetic images generated by a Generative Adversarial Network (GAN). In another research venue, studies have focused on generative OSR algorithms (Geng et al., 2020), which incorporate input reconstruction errors into the deep neural network training process to support the classification of samples (Oza & Patel, 2019; Sun et al., 2020; Yoshihashi et al., 2019). Yoshihashi et al. (2019) introduced a framework for Classification-Reconstruction learning for OSR. Oza and Patel (2019), in turn, proposed class-conditioned auto-encoders for the OSR problem, while Sun et al. (2020) focused on a scheme based on Conditional Gaussian Distribution Learning.

26.3 Applications

► Example 1 Team Tactics Estimation in Soccer Analysis

This problem consists of estimating teams' tactics based on soccer videos, often relying on players' formation on the pitch. Suzuki et al. (2018), for example, introduced an approach to establish the relationship between the tactics of two teams using the Deep-Extreme Learning Machine (DELM). Five tactics are considered (retreat, forecheck, set piece, possession, and swift attach), characterizing a closed scenario. The proposed system, therefore, does not account for tactics that differ from the pre-defined list, limiting its use in practice. ◀

► Example 2 Action Recognition for Soccer Analysis

This problem refers to the recognition of relevant actions in soccer videos. Typical actions include passing, shooting, heading, and dribbling. Ganesh et al. (2019), for example, introduced Gaussian Weighted event-based Action Classifier (GAWAC)—a Convolutional Neural Network architecture—for the problem. The proposed solution was validated on a six-class dataset that includes short pass, long pass, heading, trapping, turning, and dribbling. No “unknown” action is considered at the testing phase in their formulation, i.e., the problem is formulated as a closed-set recognition task. In real-world usage scenarios, such unknown actions will be mistakenly assigned to one of the six classes considered in the training. ◀

► Example 3 Event Recognition in Basketball Videos

This problem refers to the identification of relevant events based on spatiotemporal features extracted from basketball videos. For example, Wu et al. (2020) investigated the integration of local and global motion patterns in group activity recognition. Later, this information is combined with visual information related to predictions of success and failure (e.g., scoring) toward classifying the target event. The study also considered a closed-scenario setting encompassing six events: 3-point, free throw, layup, 2-point, slam dunk, and steal. This modelling, therefore, does not account for “unknown” events. ◀

? Questions for the Students

1. Define the OSR problem.
2. Provide two additional examples of how OSR could be used in sports analysis.
3. Pick one of the examples selected in the previous question. Which open-set recognition method described by Geng et al. (2020) could be used for the problem? Why?

Study Box

To the best of our knowledge, the investigation of OSR solutions in the context of the Sports Science domain is still an overlooked problem in the literature (refer to Boulton et al. (2019) for a list of applications of OSR). Recent initiatives have investi-

gated open-set approaches in the context of Sports data analysis. Yoon et al. (2019) investigated the use of a deep learning architecture for open-set recognition that explores spatiotemporal representation based on the learning of motion and appearance patterns. The study investigated the action recognition problem using datasets composed of actions such as running, walking, climbing, jumping, ball kicking, etc. In another study, Burns et al. (2022) investigated the use of deep triplet embeddings for personalized activity recognition using data obtained by inertial sensors. In their formulation, a deep learning model is trained by minimizing the distances of samples belonging to the same class and maximizing the distances of those of different categories. Performed validation included datasets comprising not one daily but also exercise and physiotherapy activities.

References

- Bendale, A., & Boulton, T. E. (2016). Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1563–1572).
- Boulton, T. E., Cruz, S., Dhamija, A. R., Gunther, M., Henrydoss, J., & Scheirer, W. J. (2019, July). Learning and the unknown: Surveying steps toward open world recognition. In *Proceedings of the AAAI conference on artificial intelligence* (vol. 33, no. 1, pp. 9801–9807).
- Burns, D., Boyer, P., Arrowsmith, C., & Whyne, C. (2022). Personalized activity recognition with deep triplet embeddings. *Sensors*, 22(14), 5222.
- Cardoso, D. O., Gama, J., & França, F. M. (2017). Weightless neural networks for open set recognition. *Machine Learning*, 106(9), 1547–1567.
- de Oliveira Werneck, R., Raveaux, R., Tabbone, S., & da Silva Torres, R. (2019). Learning cost function for graph classification with open-set methods. *Pattern Recognition Letters*, 128, 8–15.
- Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2021). Artificial intelligence and business value: A literature review. *Information Systems Frontiers*, 1–26.
- Ganesh, Y., Sri Teja, A., Munnangi, S. K., & Rama Murthy, G. (2019, June). A novel framework for fine grained action recognition in soccer. In *International work-conference on artificial neural networks* (pp. 137–150). Springer.
- Ge, Z., Demyanov, S., Chen, Z., & Garnavi, R. (2017). Generative openmax for multi-class open set classification. arXiv preprint arXiv:1707.07418.
- Geng, C., Huang, S. J., & Chen, S. (2020). Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10), 3614–3631.
- Goes, F. R., Meerhoff, L. A., Bueno, M. J. O., Rodrigues, D. M., Moura, F. A., Brink, M. S., et al. (2021). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, 21(4), 481–496.
- Liang, S., Li, Y., & Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690.
- Mendes Júnior, P. R., De Souza, R. M., Werneck, R. D. O., Stein, B. V., Pazinato, D. V., de Almeida, W. R., et al. (2017). Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3), 359–386.
- Naik, B. T., Hashmi, M. F., & Bokde, N. D. (2022). A comprehensive review of computer vision in sports: Open issues, future trends and research directions. *Applied Sciences*, 12(9), 4429.
- Neira, M. A. C., Júnior, P. R. M., Rocha, A., & Torres, R. D. S. (2018). Data-fusion techniques for open-set recognition problems. *IEEE Access*, 6, 21242–21265.
- Oza, P., & Patel, V. M. (2019). C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2307–2316).

- Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *Springerplus*, 5(1), 1410.
- Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., & Boulton, T. E. (2012). Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7), 1757–1772.
- Sun, X., Yang, Z., Zhang, C., Ling, K. V., & Peng, G. (2020). Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13480–13489).
- Suzuki, G., Takahashi, S., Ogawa, T., & Haseyama, M. (2018, October). Team tactics estimation in soccer videos via deep extreme learning machine based on players formation. In *2018 IEEE 7th global conference on consumer electronics (GCCE)* (pp. 116–117). IEEE.
- Wu, L., Yang, Z., Wang, Q., Jian, M., Zhao, B., Yan, J., & Chen, C. W. (2020). Fusing motion patterns and key visual information for semantic event recognition in basketball videos. *Neurocomputing*, 413, 217–229.
- Yoon, Y., Yu, J., & Jeon, M. (2019). Spatio-temporal representation matching-based open-set action recognition by joint learning of motion and appearance. *IEEE Access*, 7, 165997–166010.
- Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M., & Naemura, T. (2019). Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4016–4025).

Visualization

Contents

Chapter 27 Visualization: Basics and Concepts – 225
Daniel Link



Visualization: Basics and Concepts

Daniel Link

Contents

- 27.1 Example Sport – 226**
- 27.2 Background – 226**
- 27.3 Applications – 227**
- References – 231**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>.
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>.

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Visualizations in sports serve to visualize patterns, trends, or outliers in sports-related information, such as physiological, biochemical, or technical-tactical performance data.
- The goal is to enable a quick answer to performance diagnostic questions, for example regarding the tactical behavior of the opponent, strength and weakness profiles or correlations between training input and performance development.
- In addition to general graphical elements such as network diagrams or bar charts, it is often worthwhile to use sport-specific visualization methods that are tailored to the respective cognitive interest.

27.1 Example Sport

Visualization in sports can be found in a whole slew of application fields (Link, 2018). Media companies use graphics to enhance their reporting, professional leagues prepare information on matches visually to serve the information interest of fans, coaches and performance analysts in clubs use the graphical representation of performance data to answer questions from performance analyses to support training and competition. In science, visualizations play a major role, among other things, when statements are to be made about the performance structure of sports. In this chapter, examples are given for the visualization of match progressions and individual attacking behavior in beach volleyball. Using the sport of soccer as an example, it will be shown how passing opportunities and performance variables of free kicks can be visualized.

27.2 Background

Visualization is generally understood as the graphical representation of information (Chen et al., 2007). In sports, this mostly involves physiological, psychological, biomechanical, or technical-tactical performance indicators as well as spatial and

temporal structures and relationships between performance and contextual variables. Visualizations serve to better recognize patterns, trends and outliers and to be able to answer specific questions of sports easily and intuitively. These are mostly performance analysis questions like the tactical behavior of the opponent, strength and weakness profiles or correlations between training input and performance development. However, good visualization of sports data is an art for itself. Elaborate visualizations can be completely irrelevant, just as simple methods can deliver impressive statements. Methodically, general elements such as colors, patterns, bar charts, histograms, network diagrams or heat maps, can be used. However, in many cases, it is also worthwhile to check whether a sport-specific visualization procedure, which is optimized for the respective performance-diagnostic knowledge interest, does not offer a higher added value. Examples of this will be given in the following.

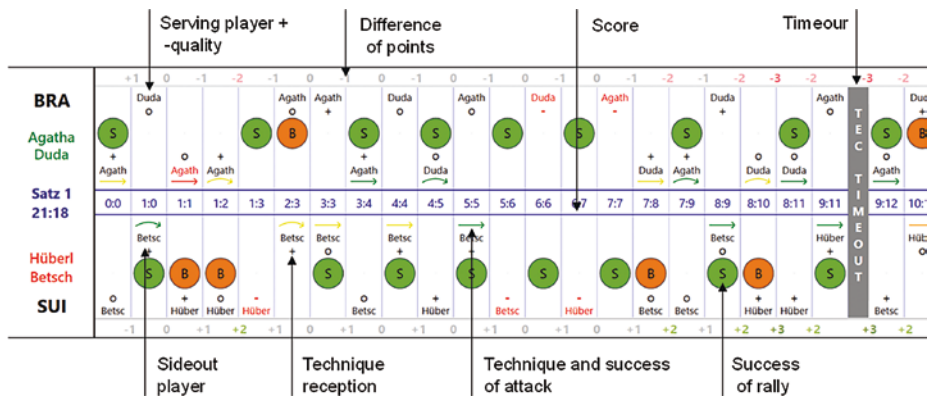
Definition

Visualization in sports is the graphical representation of sports-related information. This mostly involves physiological, psychological, biomechanical or technical-tactical performance indicators. The visualizations serve to identify patterns, trends or outliers in these performance data and to use the findings for training and competition.

27.3 Applications

► Example 1

In the German Volleyball Association, visualizations of the set structure are used in the context of opponent and self-analyses in beach volleyball. After the raw data has been collected by analysts, it is loaded into a specific analysis software (BeachViewer) (Link & Ahmann, 2013) and a graph of the match progress is generated (■ Fig. 27.1). The



■ Fig. 27.1 Illustration of a set progression in beach volleyball

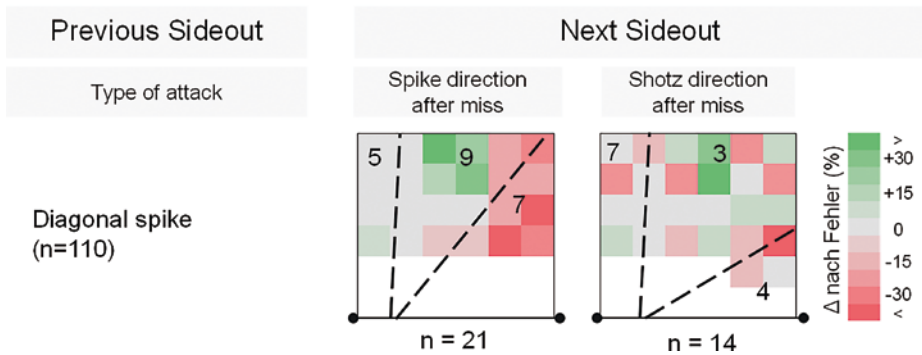
visualization is axisymmetric along the horizontal center line, which contains the score. The vertical lines separate the individual rallies and visualize timeouts.

For the first attack sequence after the serve (sideout), the temporal action sequence of a team is symbolically represented from the inside to the outside. For the serve and reception, the quality levels are indicated (–, o, +, ++), for the representation of the technique a straight (spike), curved (shot) and dashed (drive) arrow in green (point in the first attack), yellow (rally was not ended by the first attack), red (attack error or kill block) and orange (special cases) are available. The filled circle (rally success) indicates whether it is a success in the *sideout* or a *break*. Successful sideouts are marked with green and breaks with orange. The point difference gives a quick overview if and how high a team is ahead or behind. This visualization allows analysts to identify phases of strength and weakness, tactical changes or correlations between context variables (e.g. after errors, after timeouts) and player actions. This graphic is not static, but individual elements or even rallies can be hidden via filter settings. When clicking on an element, the corresponding rally is played back in the video, so that a qualitative analysis of the situation can take place. ◀

27

► Example 2

In beach volleyball—as in other sports—it is advisable to consider actions in their temporal context (Link, 2022). A relevant question in the context of strategy development is how a player acts after his own misses in the attack. Specialized representations adapted to this question make this possible. ■ Figure 27.2 shows for one player in five games the number of attacks by field zone (line, center, diagonal) for the techniques hard hit (left) and shot (right), which played after a hard-diagonal hit into the block. However, the performance diagnostic significance of the frequencies alone is not very meaningful—their value only arises when they are presented relative to a norm. For this purpose, a color coding of the target zones is used: the green coloring of a zone codes a positive deviation from the norm, a red coloring a negative one. The norm here is the distribution of the attack direction of all hard shots in the sample (i.e. not only by error).



■ Fig. 27.2 Representation of the spatial distribution of attacks in beach volleyball after misses in the previous rally by court zone. The colors code the deviation in relation to the spatial distribution of all attacks

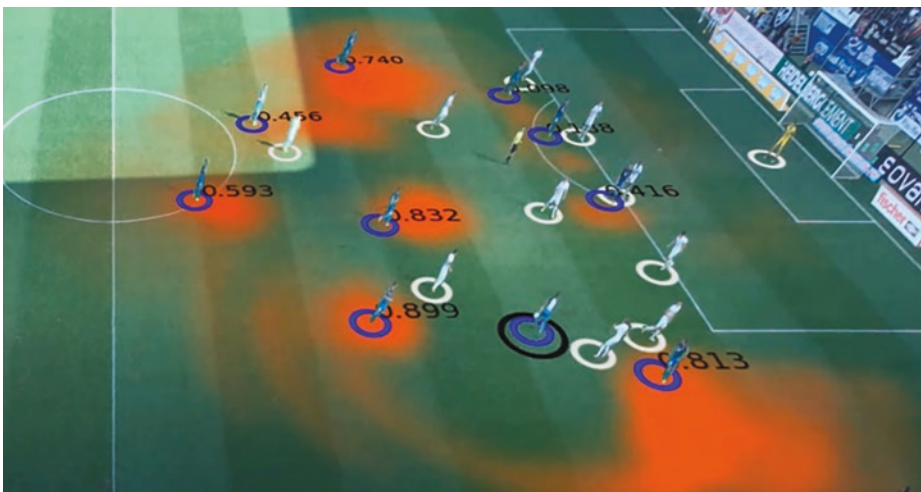
In this example, it can be seen that after a diagonal spike into the block, the player shows a tendency not to spike diagonally again, but rather tries to hit the middle area of the field in case of another hard spike. If the attack is performed as a shot, no systematic deviation from the norm can be detected visually. A possible practical consequence of this analysis could be not to make another diagonal block against this player after he has spiked diagonally into the block, but at least to try to cover the middle with one hand. ◀

► Example 3

In soccer game analyses, coaches and video analysts sometimes refer—explicitly or implicitly—to the concept of availability. This tactical construct describes the probability of success with which a player can pass the ball to a chosen teammate in a game situation. Availability is related to is related to pass risk (Power et al., 2017), but does not refer to a retrospective view of the risk of a pass played in reality; rather, it asks whether there was any possible pass at all that a teammate could reach with an acceptable chance of success and in a space that was worthwhile from a tactical perspective.

■ Figure 27.3 shows a visualization of availability for a moment in a soccer game. The black circle marks the player with possession of the ball, red areas show the availability of a player, whereby the degree of transparency, indicates the probability of a successful pass to that location. The players' labels show the accumulated probability value for a successful pass. The calculation is based on spatiotemporal data (see ► Chap. 10) and a physical model calibrated via machine learning methods using ~100,000 real passes of the soccer Bundesliga. The model is described in detail by Dick et al. (2022) and uses the duration for each player to reach an interception location on the ball trajectory based on a player motion model. The calculation is done for a variety of ball trajectories, ball velocities, ball heights, and interception points.

The application of such a visualization essentially lies in the support of the evaluation of individual game situations in the context of qualitative game analyses. For example,



■ Fig. 27.3 Visualization of the playability of potential pass receivers in soccer

situations can be automatically extracted in which only a few players were playable after winning the ball or in which unfortunate passing decisions were made. For the passer, the visualization process can be used to show possible face-off stations or to make risk-benefit trade-offs transparent. On the side of the pass receiver, the question can be answered whether players offer themselves to players in free spaces, or how long a player needs for this. Both possibly provide valuable information for training design. ◀

? Questions for the Students

1. Name the goals and fields of application for the visualization of the information in sports!
2. Give two concrete examples of application of how the visualization can be useful for sports!

Study Box

In the study “A Topography of Free Kicks in Soccer” (Link et al., 2016), the effects of their execution location on the characterizing performance variables were investigated on the basis of a sample of 1833 free kicks from the Bundesliga soccer league. Instead of simple heatmaps, so-called isomaps (Stöckl et al., 2012) were used, which continuously (and not discretely) represent the mean value of a variable on a two-dimensional surface via color gradients (■ Fig. 27.4). Using this visualization, it was possible to show for example, how centrality and proximity to the goal influence the type of execution of the free kick (■ Fig. 27.4, left). The visualization of this relationship provides a kind of norm for the goal kick vs. cross/pass decision in professional soccer. Similarly, it was shown that crosses from the right tended to be more successful than from the left (■ Fig. 27.4, right). The reason for this is that crosses from the right tend to be played by right-footed players with a trajectory away from the goal and from the left with a trajectory towards the goal. Since more players are right-footed, it can be argued that this trajectory was more successful, as the balls may have been less likely to be intercepted by the goalkeeper. Likewise, the figure shows that crosses from the back of the defense near the goal line were particularly successful. This may be due to the fact that offensive players were able to initiate running movements to the ball more quickly due to the lack of offside danger.

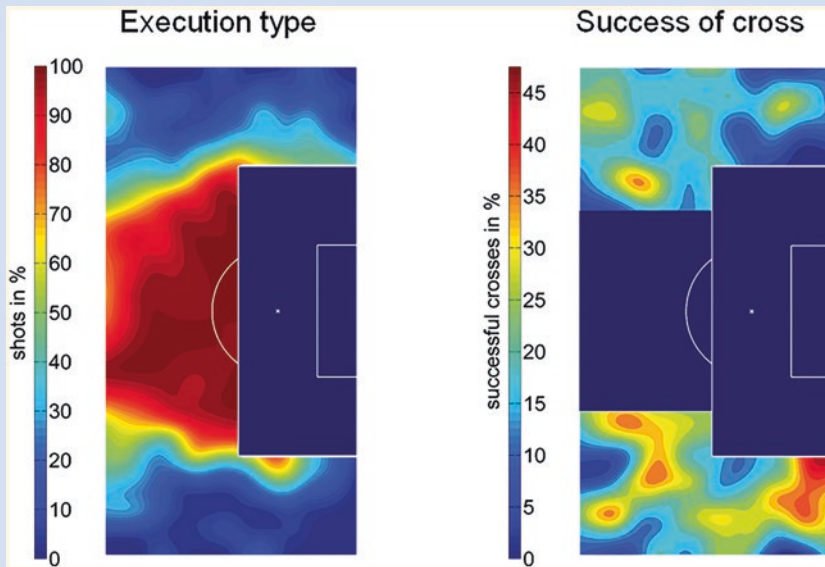


Fig. 27.4 Spatial distribution of the values of performance variables of free kicks shown via ISO maps. The left graph yields a color coding of the proportion of free kicks played as a goal kick (rather than cross or pass). The right graph represents the proportion of successful (first after ball contact by teammate) crosses

References

- C. H. Chen, W. K. Härdle, & A. Unwin (Eds.). (2007). *Handbook of data visualization*. Springer Science & Business Media.
- Dick, U., Link, D., & Brefeld, U. (2022). Who can receive the pass?—a computational model for quantifying availability in soccer. *Data Mining and Knowledge Discovery*, 36(3), 987–1014. <https://doi.org/10.1007/s10618-022-00827-2>
- Link, D. (2018). Sports analytics—how (commercial) sports data create new opportunities for sports science. *German Journal of Exercise and Sport Research*, 48(1), 13–26. <https://doi.org/10.1007/s12662-017-0487-7>
- Link, D. (2022). Spielanalyse in der Praxis: Beachvolleyball. In D. Memmert (Ed.), *Spielanalyse im Sportspiel* (pp. 43–51). Springer Spektrum Berlin, Heidelberg.
- Link, D., & Ahmann, J. (2013). Moderne Spielbeobachtung auf Basis von Positionsdaten. *Sportwissenschaft*, 43(1), 1–11. <https://doi.org/10.1007/s12662-013-0282-z>
- Link, D., Kolbinger, O., Weber, H., & Stöckl, M. (2016). A topography of free kicks in soccer. *Journal of Sports Sciences*, 34(24), 2312–2320. <https://doi.org/10.1080/02640414.2016.1232487>
- Power, P., Ruiz, H., Wei, X., Lucey, P. (2017). Not all passes are created equal. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, Halifax (Canada) 2017* (pp. 1605–1613). ACM. <https://doi.org/10.1145/3097983.3098051>.
- Stöckl, M., Lamb, P. F., & Lames, M. (2012). A model for visualizing difficulty in golf and subsequent performance rankings on the PGA tour. *International Journal of Golf Science*, 1(1), 10–24. <https://doi.org/10.1123/ijgs.1.1.10>

Outlook

Contents

Chapter 28 Outlook – 235
Arnold Baca



Outlook

Arnold Baca

Contents

- 28.1 Trends – 236**
- 28.2 Sensors – 236**
- 28.3 Wearables und Intelligent Systems – 237**
- 28.4 Big Data and Cloud – 238**
- 28.5 Machine Learning and Computer Vision – 239**
- 28.6 Virtual und Augmented Reality and Robotics – 239**
- 28.7 Data Protection and Data Misuse – 240**
- References – 240**

Digital Questions and Answers

Test your learning and check your understanding of this book's contents: use the "SpringerNature Flashcards" app to access questions. To use the app, please follow the instructions below:

1. Go to ► <https://flashcards.springernature.com/login>.
2. Create a user account by entering your e-mail address and assigning a password.
3. Use the following link to access your SN Flashcards set: ► <https://sn.pub/aWxG2v>.

If the link is missing or does not work, please send an e-mail with the subject "SN Flashcards" and the book title to customerservice@springernature.com

Key Messages

- Novel sensors will make a significant contribution to online exercise monitoring or estimation of internal body processes and external influences during physical activities.
- Data sets obtained via sensor data fusion allow the derivation of relevant physiological and tactical information, but also important conclusions about injury risks.
- The variety of technological possibilities will also lead to changes in the way sports are practiced.
- Data must be collected under clearly defined, standardized conditions so that studies can be compared.

28.1 Trends

Recent publications in the sports informatics journal *International Journal of Computer Science in Sport* and topics of papers presented at relevant conferences in 2022 (in particular the *13th International Symposium Computer Science in Sport*, which took place in Vienna from September 10–13, 2022) indicate future focal points in sports informatics in the area of the use of modern sensor technology, wearables and intelligent systems, big data and machine learning, computer vision as well as virtual and augmented reality and—to some extent—robotics. Developments and challenges in these areas are outlined below.

28.2 Sensors

In addition to established technologies (Baca, 2015) for determining type (classification) and duration (quantification) of activities (especially accelerometers), for object tracking (inertial sensors, GPS), for force measurement (e.g. strain gauges),

for estimating oxygen saturation (pulse oximeter/photo-plethysmograph) or for determining muscular activity (surface electromyograph), novel sensors are also increasingly being used that can make a significant contribution to online training control or to estimating internal body processes and external influences during physical activities. To be mentioned here are for example

- Smart textiles in which the sensor technology is integrated directly into the textile fabric. These can, for example, detect stretch, which can be used to determine breathing rate, determine humidity, temperature and other body parameters, but also react to environmental influences.
- Implantable sensors that continuously measure certain parameters (e.g. vital signs)
- Biological Biomarker Sensors for determining parameters of biological processes. Electrochemical, surface plasmon resonance-based and metamaterial-based sensor technology is promising, for example.

28.3 Wearables und Intelligent Systems

In several market studies, wearables were and are seen as a central trend in the development of sports. Areas of application include monitoring activity, tracking, promoting motivation to exercise, assessing fitness (Passos et al., 2021), and providing feedback. Recent developments (cf. Nithya & Nallavan, 2021; Lutz et al., 2019) suggest the potential of holistic approaches, where different parameter values of single or multiple exercising individuals are considered in their interplay and interaction, and of sensor data fusion, where collected data from multiple sensors are linked. It can be expected that relevant information on physiological processes, movement execution, and tactics can be derived from this objectively determined data material, but also that important conclusions on injury risks can be made possible. Challenges (Lutz et al., 2019; Rana & Mittal, 2021; Zhang et al., 2019) lie in still existing deficits in the longer-term power supply of the systems used, the availability of sensors to capture effective biomarkers to assess internal physiological processes or responses, and real-time feedback of biomechanical parameters. Future developments will increasingly focus on aspects of presentation and the way information is rendered—an optical display worn on the arm is not always the ideal solution.

Mencarini et al. (2019), based on a literature review, considered which aspects in the development of human computer interfaces (HCI) are worth considering when developing wearables in sport. In particular, it is suggested,

- Consider alternatives to the wristwatch (shape and positioning) and to simultaneous feedback—here, for example, head-up displays (HUD), where information is presented in such a way that no change in viewing direction is required, could become more important
- Place special focus on individual strengths and weaknesses and integrate support functions for groups
- To increasingly consider cognitive and emotional aspects

It can be assumed that the variety of technological possibilities will also lead to changes in the practice of sports.

Intelligent systems that shape their behavior autonomously to a certain extent and manage this depending on environmental conditions and functionality will play an even greater role in the future. Examples are the Mobile Motion Advisor (Preuschl et al., 2010), which provides feedback depending on the current performance and capacity, or intelligent strength training devices, which provide recommendations for further movement execution based on information collected during the sporting activity.

28.4 Big Data and Cloud

Experimental investigations and studies in sports are characterized by the accumulation of large amounts of data. A wide range of methods is available to manage and analyze the data. To store the data, the use of cloud storage will increase in importance. In this case, the data is stored, released and managed at a system remote from the site. In addition to issues of data integrity and data quality, where it is important that data is recorded and retrieved as intended and that unintentional changes do not affect results, as well as data availability, there will also be an even greater focus on data comparability and data completeness in the future. This means that data must be collected under clearly defined, standardized conditions and studies must thus be comparable. To this end, widespread international collaboration on aggregation and harmonization of “open access” datasets should be sought (Richter et al., 2021). Phatak et al. (2021) lament a deficit in the collection of potentially extensive, high-quality “Big Data” in an organized, time-synchronized, and holistic manner.

Big Data in sports offers significant opportunities in collaborations between sports science and computer science, such as in tactics analysis in soccer, especially in the use of positional data (Goes et al., 2021).

There is also great potential in the analysis of data collected on various fitness platforms. Here, too, poor data quality is still a limiting factor. For example, Zrenner et al. (2021) conducted a retrospective analysis of training and its effects in marathon runners using data from fitness apps. This did not guarantee that all subjects logged and uploaded all physical activities. Similarly, contextual information affecting performance, such as humidity and temperature during a workout or injury, was not available.

In addition, the analysis of sports-related data from social media is also in its infancy.

In order to integrate data from a wide variety of measurement and information systems, sports organizations, associations and clubs are increasingly requiring, using and maintaining sports and club information systems. Here, too, there is a need for generalized and standardized concepts (Blobel & Lames, 2020; Blobel et al., 2021).

28.5 Machine Learning and Computer Vision

The availability of big data and the potential of machine learning methods have also led to the rapidly increasing use of data-based approaches for different questions in sports and sports science (Bai & Bai, 2021). For team sports, for example, the following application areas can be distinguished (cf. Baca, 2021):

- Result prediction (Horvat & Job, 2020)
- Injury prevention (Van Eetvelde et al., 2021) & Health monitoring (Wu et al., 2021).
- Analysis of the performance of players
- Recognition of movement patterns
- Tactics and strategy analysis

This development will continue in the coming years. This will be particularly true for the prediction of match and competition results, as large sums of money are involved in the betting sector. Furthermore, an increased use for enriching sports broadcasts with additional information is to be expected.

However, machine learning methods have also been used for a long time in the field of computer vision or for understanding and interpreting digital images and videos. With the development of deep learning-based methods, impressive results can be achieved here that also open up previously unknown possibilities for sports. By automatically identifying two- or three-dimensional structures and features from individual two-dimensional images or series of such images taken by a single camera or by multiple cameras, information can be extracted both on movements of individual persons or their body segments and on interacting teams.

28.6 Virtual und Augmented Reality and Robotics

To support training and expand the presentation options for viewers, further impressive developments can also be expected in the field of Virtual (VR) and Augmented (also Enhanced; AR) Reality. Whereas in VR people act interactively in a virtual environment generated in real time, in AR real objects and environments are also displayed three-dimensionally.

In training, this particularly concerns virtual and augmented environments for interaction with virtual characters as preparation for competitions (Petri et al., 2018). In this way, for example, reaction skills can be trained in the sport of karate. In the future, in addition to optical sensory impressions, acoustic (e.g., the immersion of a rowing blade in water) and haptic (e.g., touching sports equipment) feedback will increasingly be simulated.

In the media sector, the aim is to create opportunities to experience sporting events directly from the point of view of the player in question and thus really be in the middle of the action.

Probably the best-known example of the use of robots in sports is robot soccer. This field of application will probably continue to be a test scenario for the use of new methods of artificial intelligence and sensor technologies, since the goal in RoboCup is still to beat the soccer world champion by the year 2050.

However, other possible applications are also being advanced. For example, humanoid robots could also be used to guide sports exercises and support (fitness) trainers (Griffiths et al., 2021).

28.7 Data Protection and Data Misuse

Many of the methods and technologies addressed in this section collect, process or store confidential, personal data. Protection of this data is not always guaranteed. Fitness data collected via apps and wearables, for example, is often uploaded directly to clouds or transmitted to providers. This can certainly be associated with risks of which one should be aware.

Questions for the Students

1. Name open challenges in the development of wearables for sports.
2. What developments can be expected in the field of virtual reality in sports?

References

- Baca, A. (2015). Data acquisition and processing. In A. Baca (Ed.), *Computer science in sport: Research and practice* (pp. 46–81). Routledge.
- Baca, A. (2021). Machine learning. In J. Pino-Ortega & M. Rico-Gonzalez (Eds.), *The use of applied technology in team sport* (pp. 230–241). Routledge.
- Bai, Z., & Bai, X. (2021). Sports big data: Management, analysis, applications, and challenges. *Complexity*, 2021, 6676297. <https://doi.org/10.1155/2021/6676297>
- Blobel, T., & Lames, M. (2020). A concept for club information systems (CIS)—an example for applied sports informatics. *International Journal of Computer Science in Sport*, 19(1), 102–122. <https://doi.org/10.2478/ijcss-2020-0006>
- Blobel, T., Rumo, M., & Lames, M. (2021). Sports information systems: A systematic review. *International Journal of Computer Science in Sport*, 20(1), 1–22. <https://doi.org/10.2478/ijcss-2021-0001>
- Goes, F. R., Meerhoff, L. A., Bueno, M. J. O., Rodrigues, D. M., Moura, F. A., Brink, M. S., Elferink-Gemser, M. T., Knobbe, A. J., Cunha, S. A., Torres, R. S., & Lemmink, K. A. P. M. (2021). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, 21, 481–496. <https://doi.org/10.1080/17461391.2020.1747552>
- Griffiths, S., Alpay, T., Sutherland, A., Kerzel, M., Eppe, M., Strahl, E., & Wermter, S. (2021). Exercise with social robots: Companion or coach? arXiv:2103.12940 [cs]. <https://arxiv.org/abs/2103.12940v1>.
- Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *WIREs Data Mining and Knowledge Discovery*, 10(5), e1380. <https://doi.org/10.1002/widm.1380>

- Lutz, J., Memmert, D., Raabe, D., Dornberger, R., & Donath, L. (2019). Wearables for integrative performance and tactic analyses: Opportunities, challenges, and future directions. *International Journal of Environmental Research and Public Health*, 17(1), 59. <https://doi.org/10.3390/ijerph17010059>
- Mencarini, E., Rapp, A., Tirabeni, L., & Zancanaro, M. (2019). Designing Wearable Systems for Sports: A Review of Trends and Opportunities in Human–Computer Interaction. *IEEE Transactions on Human-Machine Systems*, 49(4), 314–325. <https://doi.org/10.1109/THMS.2019.2919702>.
- Nithya, N., & Nallavan, G. (2021). Role of wearables in sports based on activity recognition and biometric parameters: A survey. In *2021 international conference on artificial intelligence and smart systems (ICAIS)*. <https://doi.org/10.1109/icaais50930.2021.9395761>.
- Passos, J., Lopes, S. I., Clemente, F. M., Moreira, P. M., Rico-González, M., Bezerra, P., & Rodrigues, L. P. (2021). Wearables and internet of things (IoT) technologies for fitness assessment: a systematic review. *Sensors (Basel)*, 21(16), 5418. <https://doi.org/10.3390/s21165418>
- Petri, K., Bandow, N., & Witte, K. (2018). Using several types of virtual characters in sports—a literature survey. *International Journal of Computer Science in Sport*, 17(1), 1–48. <https://doi.org/10.2478/ijcss-2018-0001>
- Phatak, A. A., Wieland, F.-G., Vempala, K., Volkmar, F., & Memmert, D. (2021). Artificial intelligence based body sensor network framework—narrative review: Proposing an end-to-end framework using wearable sensors, real-time location systems and artificial intelligence/machine learning algorithms for data collection, data mining and knowledge discovery in sports and healthcare. *Sports Medicine—Open*, 7(1), 79. <https://doi.org/10.1186/s40798-021-00372-0>
- Preuschl, E., Baca, A., Novatchkov, H., Kornfeind, P., Bichler, S., & Boecksoer, M. (2010). Mobile motion advisor—A feedback system for physical exercise in schools. *Procedia Engineering*, 2(2), 2741–2747. <https://doi.org/10.1016/j.proeng.2010.04.060>
- Rana, M., & Mittal, V. (2021). Wearable sensors for real-time kinematics analysis in sports: A review. *IEEE Sensors Journal*, 21(2), 1187–1207. <https://doi.org/10.1109/jsen.2020.3019016>
- Richter, C., O'Reilly, M., & Delahunt, E. (2021). Machine learning in sports science: Challenges and opportunities. *Sports Biomechanics*, 81, 1–11. <https://doi.org/10.1080/14763141.2021.1910334>
- Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. *Journal of Experimental Orthopaedics*, 8(1), 27. <https://doi.org/10.1186/s40634-021-00346-x>
- Wu, X., Liu, C., Wang, L., & Bilal, M. (2021). Internet of things-enabled real-time health monitoring system using deep learning. *Neural Computing and Applications*, 35, 14565–14576. <https://doi.org/10.1007/s00521-021-06440-6>
- Zhang, X., Shan, G., Wang, Y., Wan, B., & Li, H. (2019). Wearables, biomechanical feedback, and human motor-skills' learning & optimization. *Applied Sciences*, 9(2), 226. <https://doi.org/10.3390/app9020226>
- Zrenner, M., Heyde, C., Duemler, B., Dykman, S., Roecker, K., & Eskofier, B. M. (2021). Retrospective analysis of training and its response in Marathon finishers based on fitness app data. *Frontiers in Physiology*, 12, 669884. <https://doi.org/10.3389/fphys.2021.669884>

Supplementary Information

Appendix. Third-Party Funds Competitively
Acquired by German Sports Scientists from
the German Research Foundation (DFG) in the
Review Board for Computer Science – 244

Index – 247

Appendix. Third-Party Funds Competitively Acquired by German Sports Scientists from the German Research Foundation (DFG) in the Review Board for Computer Science

| Applicant | Title | Funding period | Link |
|----------------|--|----------------|---|
| Daniel Memmert | Implementation of floodlight e-Research technology for the analysis of spatio-temporal motion data in sports science | 2023–2026 | ► https://www.dshs-koeln.de/aktuelles/meldungen-pressemittelungen/detail/meldung/sportspieldaten-effektiv-nutzbar-machen/ |
| Kerstin Witte | Visual peripheral perception in virtual reality | 2023–2025 | ► https://gepris.dfg.de/gepris/projekt/404484468 |
| Daniel Memmert | Data-based approaches to the analysis of soccer matches from an e-science perspective | Since 2020 | ► DFG—GEPRIS—Data-based approaches to the analysis of soccer matches from an e-science perspective |
| Daniel Memmert | A theoretical simulation framework for the analysis of predictive rating methods on networks with applications in sports | 2019–2025 | ► https://gepris.dfg.de/gepris/projekt/432919559?context=projekt&task=showDetail&id=432919559& |
| Kerstin Witte | Training in VR with special emphasis on visual perception and comparison to reality | 2018–2022 | ► https://gepris.dfg.de/gepris/projekt/404484468 |
| Daniel Memmert | Simulation of interactive action sequences using the example of high-performance soccer | 2018–2025 | ► DFG—GEPRIS—Simulation of interactive action sequences using the example of high-performance soccer |
| Dietmar Saupe | Powerbike—Model-based optimization for road cycling | 2013–2018 | ► https://gepris.dfg.de/gepris/projekt/432919559?context=projekt&task=showDetail&id=432919559& |

| | | | |
|-------------------------------|--|-----------|---|
| Kerstin Witte
G. Brunnett | Development of an autonomously interacting opponent in a virtual reality environment to study anticipation ability in martial arts | 2014–2016 | ▶ https://gepris.dfg.de/gepris/projekt/252070407 |
| Daniel Memmert
Jürgen Perl | Simulation of interaction patterns and simulative effectiveness analysis of creative actions in sports game using neural networks | 2008–2018 | ▶ DFG—GEPRIS—Simulation of Interaction Patterns and Simulative Effectiveness Analysis of Creative Actions in Sports Games Using Neural Networks |

Index

A

Acceleration, 77
Actions, 28
Activation functions, 172
AI-based approaches, 28
Analysis routines, 127
Antagonism, 101
Antagonistic models, 104
Application Programming Interface (API), 51
Artificial data, 16
Artificial intelligence, 202
Artificial neural networks, 170, 172
Athletic performance, 100
Augmented reality, 239–240
Automated notation, 38
Automatic text classifiers, 24

B

Beach volleyball, 226
Benchmark datasets, 178, 179, 183
Big Data, 137, 238
Binary outcome, 136
Biomarker sensors, 237

C

Calculation, 58
Carbon dioxide output, 90
Classification algorithms, 136
Cloud storage, 238
Clustering, 144
Coaching, 143
Command line, 113
Complex dynamic systems, 58
Complex processes, 14
Complex system, 39
Computer science, 194
Computer vision, 8, 178, 179, 218, 239
Computing systems, 194
Conformance checking, 152
Context information, 47
Contextual factors, 47

Convolutional neural networks (CNNs), 179, 181, 186
CRAN, 112
Critical power, 92

D

Data analysis, 110
Data availability, 15, 66
Data fusion, 236
Data imbalances, 137
Data mining, 142
Data processing pipelines, 127
Data representation, 194
Decision making, 137, 142
Decision tree, 202
Deep learning, 29, 180
Definition, 4
Differential equations, 92
Distance, 77
Domain-independent, 195
Dropout, 179
dvs section Sportinformatik, 7
Dynamic pricing, 50

E

Efficiency, 100
ELO rating, 69
Energy provision, 90
Ensemble methods, 202
Error function, 172
Event data, 36, 142
Expected goals, 37, 186
External load, 74

F

Fatigue, 100
Feature vector, 195
Features, 179, 182, 187
Field registration, 28
Forecast, 175
Fully-connected neural networks, 187

G

Genders, 46
 Gradient boosting, 210
 Graph-based neural networks, 182

H

HTML, 51
 Hydraulic model, 95

I

IACSS, 7
 Image processing, 194
 Indicators, 45
 Individual case, 101
 Information, 195
 Institutionalization, 4
 Interactive behavior, 158
 Interactive programming, 128
 Internal load, 74

K

Key performance indicators (KPIs), 45, 182, 203
 K-nearest neighbor, 181
 Knowledge-discovery process, 137

L

Labels, 203
 Lactate, 90
 Lasso, 211
 Learning rules, 172
 Lexicon-based categorization of text data, 25
 Libraries, 128
 Literate programming, 121
 Live-betting odds, 53
 Load, 100
 Logistic regression model, 69

M

Machine learning, 8, 66, 218, 229, 239
 Manual scraping, 51
 Mapping, 58
 Markdown, 121
 Market efficiency, 67
 Medicine, 145
 Metabolic power, 77

Metadata, 28
 Modeling, 58, 82, 210
 Model parameters, 102
 Monte Carlo Simulation, 16
 Movements, 28
 Multimedia data, 218

N

Network analysis, 158
 Network science, 158
 Neural networks, 46, 187
 Norm, 229

O

Online data, 50
 Open-set recognition, 218
 Open-source, 128
 Oxygen uptake, 90

P

Packages, 112
 Passing networks, 39
 Pattern recognition, 174, 218
 Performance analysis, 82, 158
 PerPot model, 83
 Physiological models, 74
 Physiological performance, 85
 Positional data, 44, 142, 186
 Poisson distribution, 211
 Prediction, 210
 Prediction of performance, 143
 Predictive models, 66
 Predictive quality, 66
 Process discovery, 151
 Process enhancement, 152
 Process models, 151
 Profiles, 84
 Programming languages, 126
 Propagation function, 171

R

Random forests, 69, 202, 210
 Random numbers, 14
 Recovery, 100
 Recurrent neural networks, 179
 Regression, 210
 Regression analysis, 94
 Representation, 187

RMarkdown, 121
Robotics, 239–240
R programming language, 110

S

Scientific discipline, 4
Sentiment analysis, 23
Sequence, 37
Sequence analysis, 40
Simulation, 82
Soccer, 28, 46, 226
Soccer matches, 210
Social Network Analysis, 159
Spatiotemporal data, 229
Speed zones, 76
Sports betting market, 67
Sports informatics, 4
Statistical/machine learning, 210
Statistical modelling, 17, 66
Strain, 100
Supervised, 170
Supervised statistical learning, 211
System reduction, 59

T

Tactical, 85

Text mining, 22
Time series, 142
Training effect analyses, 104
Transfer learning allows, 194
Transformation, 58, 195
TRIMP, 76
TSDM, 143

U

Unsupervised, 170

V

VAEP framework, 39
Velocity, 77
Video, 28
Virtual, 239–240
Visualizations, 226
Voronoi, 61

W

Wearables, 237–238
Web crawling, 50
Web scraping, 50
Wisdom of the Crowd, 24
Workflows, 151