

# Contents

|          |                                                                                         |          |
|----------|-----------------------------------------------------------------------------------------|----------|
| <b>1</b> | <b>Introduction</b>                                                                     | <b>1</b> |
| <b>2</b> | <b>Correspondence Analysis</b>                                                          | <b>4</b> |
| 2.1      | Introduction . . . . .                                                                  | 4        |
| 2.2      | Simple Correspondence Analysis . . . . .                                                | 8        |
| 2.2.1    | Notation and Computation . . . . .                                                      | 8        |
| 2.2.2    | Criteria . . . . .                                                                      | 9        |
| 2.2.3    | Extensions . . . . .                                                                    | 10       |
| 2.2.3.1  | Supplementary Points . . . . .                                                          | 10       |
| 2.2.3.2  | Correspondence Analysis of Subsets . . . . .                                            | 11       |
| 2.2.4    | Implementation of Simple Correspondence Analysis in the ca package . . . . .            | 11       |
| 2.3      | Multiple Correspondence Analysis . . . . .                                              | 14       |
| 2.3.1    | Notation and Computation . . . . .                                                      | 14       |
| 2.3.1.1  | Computation of Multiple Correspondence Analysis based on the Indicator Matrix . . . . . | 17       |
| 2.3.1.2  | Computation of Multiple Correspondence Analysis based on the Burt Matrix . . . . .      | 17       |
| 2.3.2    | Adjustment of Inertias . . . . .                                                        | 18       |
| 2.3.3    | Joint Correspondence Analysis . . . . .                                                 | 18       |
| 2.3.4    | Extensions . . . . .                                                                    | 21       |
| 2.3.4.1  | Supplementary Points . . . . .                                                          | 21       |
| 2.3.4.2  | Multiple Correspondence Analysis in the Case of Subsets . . . . .                       | 22       |

|          |                                                                                        |           |
|----------|----------------------------------------------------------------------------------------|-----------|
| 2.3.5    | Implementation of Multiple Correspondence Analysis in the ca package . . . . .         | 23        |
| 2.4      | Visualization of Results from Simple and Multiple Correspondence Analysis . . . . .    | 25        |
| 2.5      | Further Topics in Correspondence Analysis . . . . .                                    | 29        |
| 2.5.1    | Assessing the Stability of Results obtained from Correspondence Analysis . . . . .     | 29        |
| 2.5.2    | The Application of Clustering based on Correspondence Analysis . . . . .               | 31        |
| 2.6      | Existing Implementations of Simple and Multiple Correspondence Analysis in R . . . . . | 33        |
| <b>3</b> | <b>Web Mining</b> . . . . .                                                            | <b>35</b> |
| 3.1      | Introduction . . . . .                                                                 | 35        |
| 3.2      | Technical Overview . . . . .                                                           | 35        |
| 3.2.1    | The Structure of the Web . . . . .                                                     | 35        |
| 3.2.1.1  | Web Resources . . . . .                                                                | 36        |
| 3.2.1.2  | Web Resource Identifiers . . . . .                                                     | 37        |
| 3.2.1.3  | Web Transfer Protocols . . . . .                                                       | 38        |
| 3.2.2    | Data Sources for Web Mining . . . . .                                                  | 38        |
| 3.2.2.1  | The Common Log Format . . . . .                                                        | 39        |
| 3.2.2.2  | Issues in Data Retrieval . . . . .                                                     | 42        |
| 3.3      | Web Mining . . . . .                                                                   | 43        |
| 3.3.1    | Aspects in Web Mining . . . . .                                                        | 43        |
| 3.3.2    | Further Issues in Web Mining . . . . .                                                 | 44        |
| 3.3.2.1  | The Relation between Web Mining, Data Mining and Statistics . . . . .                  | 44        |
| 3.3.2.2  | Web Usage Data Abstractions . . . . .                                                  | 44        |
| 3.3.2.3  | Preprocessing Web Logs . . . . .                                                       | 45        |
| 3.4      | Selected Approaches to Web Usage Mining . . . . .                                      | 47        |
| 3.4.1    | Early Studies on Browsing Behaviour . . . . .                                          | 47        |
| 3.4.2    | Modelling Web Usage . . . . .                                                          | 48        |

|          |                                                                             |           |
|----------|-----------------------------------------------------------------------------|-----------|
| 3.4.2.1  | Representing Web Site Visits with Graphs . . . . .                          | 48        |
| 3.4.2.2  | Markov Chains . . . . .                                                     | 49        |
| 3.4.2.3  | Application of Markov Chains in the Analysis<br>of Web Usage . . . . .      | 50        |
| <b>4</b> | <b>A Case Study</b>                                                         | <b>53</b> |
| 4.1      | Introduction . . . . .                                                      | 53        |
| 4.1.1    | Web Site Structure . . . . .                                                | 54        |
| 4.1.2    | Data Preprocessing . . . . .                                                | 56        |
| 4.2      | Preliminary Results . . . . .                                               | 57        |
| 4.2.1    | Distributional Properties of Web Site Accesses . . . . .                    | 57        |
| 4.2.2    | Temporal Properties of Web Site Accesses . . . . .                          | 61        |
| 4.2.3    | Transitional Properties of Web Site Accesses . . . . .                      | 65        |
| 4.2.4    | Consideration of Target Variables . . . . .                                 | 68        |
| 4.3      | The Application of Correspondence Analysis in Web Usage<br>Mining . . . . . | 71        |
| 4.3.1    | An Analysis of Users' Web Page Accesses . . . . .                           | 71        |
| 4.3.2    | Application of Supplementary Points in Correspondence<br>Analysis . . . . . | 76        |
| 4.4      | Further Applications based on Correspondence Analysis . . . . .             | 79        |
| 4.4.1    | Clustering of Web Usage Data based on Correspondence<br>Analysis . . . . .  | 79        |
| 4.4.2    | Real Time Monitoring of Web Site Accesses . . . . .                         | 81        |
| <b>5</b> | <b>Conclusions and Future Work</b>                                          | <b>83</b> |
| <b>A</b> | <b>Correspondence Analysis in R: the ca package</b>                         | <b>85</b> |
| A.1      | Package Overview . . . . .                                                  | 85        |
| A.2      | Function Reference . . . . .                                                | 86        |
| A.2.1    | Simple Correspondence Analysis . . . . .                                    | 86        |
| A.2.2    | Multiple and Joint Correspondence Analysis . . . . .                        | 96        |
| A.2.3    | Data sets and Miscellaneous Functions . . . . .                             | 106       |

---

|          |                                                     |            |
|----------|-----------------------------------------------------|------------|
| <b>B</b> | <b>Web Mining in R: the <code>wm</code> package</b> | <b>109</b> |
| B.1      | Package Overview . . . . .                          | 109        |
| B.2      | Function Reference . . . . .                        | 110        |
| B.2.1    | Data Import and Preprocessing . . . . .             | 110        |
| B.2.2    | Visualization . . . . .                             | 116        |
| B.2.3    | Miscellaneous Functions . . . . .                   | 117        |