STATISTICAL DATA ANALYSIS AND NATURAL LANGUAGE PROCESSING FOR NUTRITION SCIENCE

Tome Eftimov

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Asst. Prof. Dr. Barbara Koroušić Seljak, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Nada Lavrač, Chair, Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Dr. Peter Korošec, Member, Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Dr. Sophia Ananiadou, Member, University of Manchester, Manchester, United Kingdom

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL



Tome Eftimov

STATISTICAL DATA ANALYSIS AND NATURAL LANGUAGE PROCESSING FOR NUTRITION SCIENCE

Doctoral Dissertation

STATISTIČNA ANALIZA PODATKOV IN OBDELAVA NARAVNEGA JEZIKA ZA PREHRANSKO ZNANOST

Doktorska disertacija

Supervisor: Asst. Prof. Dr. Barbara Koroušić Seljak

Ljubljana, Slovenia, January 2018

To my Mother, Father, Brother, Both Grandmothers and Both Grandfathers

Acknowledgments

I would like to express my special appreciation and thanks to my supervisor Asst. Prof. Dr. Barbara Koroušić Seljak. I would like to thank her for encouraging my research and for allowing me to grow as a scientist. I am especially thankful for all our discussions, not related only to this work, for giving me total freedom to be myself, and teaching me that a true accomplishment can be achieved only with something we love. Additionally, I would like to thank Peter Korošec for all his constructive feedback during this work, for our endless discussions about nothing and everything, and for teaching me that sometimes we win, but sometimes we learn. I would also like to thank the other two members of the evaluation board: Prof. Dr. Nada Lavrač and Prof. Dr. Sophia Ananiadou for the constructive feedback on this work.

I would like to thank the financial support received through the postgraduate scholarship program of the EU project, "Era Chair for isotope techniques in food quality, safety and traceability" and the Computer Systems Department at the Jožef Stefan Institute.

I am thankful to all my colleagues at the Computer Systems Department, for a really friendly and welcoming environment. Special thanks to Jolanda Jakofčič for helping me with all the administration work at the Jožef Stefan Institute during my PhD studies. I would also like to thank the ERA Chair Holder, David Heath, for English proofreading of my papers.

I am grateful to all my friends for their support, confidence and patience. First, special thanks to Hristijan Gjoreski and Martin Gjoreski for allowing me to understand and to be understood. I also appreciate all the trips spent together that resulted in some memories that are unforgettable, remaining ever vivid and heartwarming. Further, I would like to thank Ivana Kostadinovska and Velice Grčeva, or as I call them my sisters, for always being there for me and supporting my decisions unconditionally. Then, special thanks also to Gordana Gjorgieva and Tanja Ivanovska for always being prepared to hear me and for our numerous calls and hours spent talking. Thank you to Barbara Volarič and Biljana Stojković for our endless conversations that helped me learn Slovene. And let me not forget my little ones Mimi Ljoncheva, Gordana Ispirova, and Gjorgi Peev, who were always with me regardless of good or bad. I am happy to call all of you my friends, or my second family.

Finally, I am most grateful to my family for their unconditional support. Especially, I wish to express my gratitude and thankfulness to my mother, Lenče Eftimova, father, Risto Eftimov, brother, Darko Eftimov, both grandmothers, Ljuba Eftimova and Dušana Janeva, and both grandfathers, Tome Eftimov and Boris Janev. None of this would have been possible without their unconditional love and confidence in me.

At the end, this is no real ending. It is just the place where we stop one story, but we begin with another one.

Abstract

The work presented in this thesis is a synergy of statistics and natural language processing. Statistics is a branch of mathematics that is related to the collection, analysis, interpretation or explanation, and presentation of data. Working with experimental data, researchers should perform a statistical analysis to correctly interpret the results of the study, which are usually presented in scientific publications. To follow the new knowledge in a particular domain, which is quickly increasing with recently published electronic publications, computer-based methods are extremely welcome for knowledge identification, extraction and exploration. This is a well-known task in natural language processing. It is a research area of computer science, artificial intelligence, and computational linguistics, concerned with the interactions between computers and human natural languages.

In this thesis, we propose an approach to explore a given domain. The part of statistics is focused on how to obtain more robust statistical results that need to be published, and the part of natural language processing is focused on the extraction and normalization of relevant scientifically published information in order to follow the new knowledge of the domain.

In the area of statistics, we present a novel method for making a statistical comparison of experimental data that is more robust to outliers and small differences that can exist between data values. The main contribution of the method is the new ranking scheme, which uses the whole distribution of the data instead of using only one statistic to describe the data distribution, such as average or median.

In the area of natural language processing, we address the problem of information extraction from an untapped domain. We propose a new rule-based named-entity recognition method. The method does not require an annotated corpus, and the main difference with the other rule-based named-entity recognition methods is that the rules are not associated with the characteristics of the entities of interest. We improve the string similarity of domain-specific short segments of text by probability modeling of the domain using the morphological information presented in the text. The proposed method can be a basis for text normalization methods, which are methods used for automatic mapping of a concept (entity) to a concept that exists in a domain-specific terminological resource.

The methods are evaluated using real-life problems taken from computer science and nutrition science. The statistical method is tested using experimental data from the field of optimization and the evaluation results show that it gives more robust results than the commonly-used statistical comparison approach, especially when the results are affected by the presence of outliers or small differences that may exist between the data values. The rule-based named-entity recognition method is tested in the dietary domain, which is an untapped domain, from where promising results are achieved. Finally, to collect the information for the same entity that can be represented in different ways, using phrases with a variety of structures, the method for string similarity of domain-specific short segments of text is applied on food concepts, or the food matching problem. The evaluation results show that it gives more promising results compared it to commonly-used string similarity

measures applied on the same problem.

Povzetek

V disertaciji predstavljeno delo je sinergija statistike in (računalniške) obdelave naravnega jezika. Statistika je veja matematike, ki obravnava zbiranje, analizo, interpretacijo ali razlago, in predstavitev podatkov. Raziskava z eksperimentalnimi podatki mora vključevati statistično analizo potrebno za pravilno interpretacijo rezultatov, ki jih želimo predstaviti v znanstvenih publikacijah. Za spremljanje novega znanja znotraj določene domene, to se hitro povečuje s sprotnim objavljanjem elektronskih publikacij, so izjemno dobrodošle računalniško podprte metode prepoznavanja, pridobivanja, in preiskovanja ali izkoriščanja znanja. To so dobro znane naloge pri obdelavi naravnega jezika – raziskovalnega področja računalništva, umetne inteligence in računalniškega jezikoslovja, ki se ukvarja z interakcijami med računalnikom in človeškimi naravnimi jeziki.

V disertaciji predlagamo nov pristop za raziskovanje določene domene. Del disertacije, ki obravnava statistiko, je osredotočen na pridobivanje robustnejših statističnih rezultatov, ki jih je potrebno objaviti, dočim je del disertacije, ki obravnava obdelavo naravnega jezika, osredotočen na pridobivanje in normalizacijo ustreznih znanstveno relevantnih podatkov, s katerimi bo lažje spremljati nova znanja o tej domeni.

Na področju statistike predlagamo novo metodo za statistično primerjavo eksperimentalnih podatkov, ki je bolj robustna na osamelce in morebitne majhne razlike med vrednostmi podatkov. Metoda temelji na novi razvrstitveni shemi, ki uporablja celotno porazdelitev podatkov in ne le posamezne statistike za opis porazdelitve podatkov, kot sta povprečje ali mediana.

Na področju obdelave naravnega jezika obravnavamo problem pridobivanja podatkov iz do zdaj še neobravnavane domene, kjer predlagamo novo metodo prepoznanja entitet, ki temelji na pravilih. Metoda ne zahteva označenega korpusa in se od drugih metod prepoznanja entitet, ki temeljijo na pravilih, razlikuje v tem, da pravila niso povezana z značilnostmi entitet. Z uporabo verjetnostnega modela domene, ki vsebuje morfološke informacije najdene v besedilu, smo izboljšali zaznavanje podobnosti med kratkimi besedili, ki so del iste domene. Predlagana metoda je lahko osnova za normalizacijo besedila, ki je uporabljena pri avtomatski preslikavi koncepta (entitete) v koncept, ki že obstaja v terminološkem viru, specifičnem za določeno domeno.

Predlagane metode so ovrednotene na realnih problemih s področij računalništva in prehrane. Statistična metoda je preizkušena z uporabo eksperimentalnih podatkov s področja optimizacije. Rezultati kažejo, da daje metoda bolj robustne rezultate od pogosto uporabljenega pristopa za statistično primerjavo, še posebej, če na rezultate vpliva prisotnost osamelcev ali morebitnih majhnih razlik med vrednostmi podatkov. Metoda prepoznanja entitet je testirana na področju prehrane, do zdaj še neobravnavane domene, kjer so doseženi obetavni rezultati. Metoda je za iskanje podobnosti med kratkimi besedili testirana tudi na problemu ujemanja informacij o živilih, kjer je entiteta lahko predstavljena na različne načine, z uporabo besednih zvez z različnimi strukturami. Rezultati kažejo, da dosega metoda obetavnejše rezultate v primerjavi z običajnimi pristopi, ki se uporabljajo za ugotavljanje podobnosti med besedili.

Contents

List of Figures **xv**

List of Tables **xvii**

List of Algorithms **xix**

Abbreviations **xxi**

1 Introduction **1**
1.1 Problem Definition . . . 1
1.2 Hypotheses . . . 4
1.3 Scientific Contributions . . . 4
1.4 Thesis Structure . . . 7

2 Background **9**
2.1 Statistical Comparison . . . 9
2.1.1 Basic terms . . . 9
2.1.2 Hypothesis testing . . . 11
2.2 Information Extraction . . . 14
2.2.1 Natural language processing . . . 14
2.2.2 Machine learning . . . 15
2.2.3 Named-entity recognition . . . 16
2.2.4 Text normalization . . . 16

3 Deep Statistical Comparison **19**
3.1 Problem Definition . . . 19
3.2 Related Work . . . 21
3.3 Methodology . . . 22
3.3.1 The DSC ranking scheme . . . 22
3.3.2 Omnibus statistical tests . . . 26
3.3.3 The difference between practical and statistical significance in data analysis . . . 26
3.4 Experimental Results . . . 27
3.4.1 Multiple comparisons among three data sets . . . 28
3.4.2 Multiple comparisons among four data sets . . . 40
3.4.3 Multiple comparisons with a control data set . . . 40
3.4.4 Multiple comparisons among three data sets by using the sequential DSC ranking scheme . . . 47
3.4.5 Multiple comparisons among three data sets by using the Monte-Carlo DSC ranking scheme . . . 49
3.4.6 All possible pairwise comparisons . . . 51
3.5 Power Analysis . . . 56

3.6 DSC Ranking Scheme With Different Criteria of Comparing Distributions . 59
3.7 Discussion and Conclusion on DSC 63

4 Rule-Based Named-Entity Recognition 67
4.1 Problem Definition 67
4.2 Related Work 68
4.3 Methodology 70
4.3.1 Pre-processing 71
4.3.2 Detection and determination of entities mentions 72
4.3.3 Selection and extraction of the entities 74
4.4 Experimental Results 75
4.4.1 Dictionaries 76
4.4.2 Examples 77
4.4.3 Evaluation 84
4.5 Discussion and Conclusion on drNER 88

5 String Similarity of Domain-Specific Short Text Segments 93
5.1 Problem Definition 93
5.2 Related Work 94
5.3 Methodology 95
5.3.1 Part-of-speech tagging probability weighted method 95
5.4 Experimental Results 97
5.4.1 FoodEx2 97
5.4.2 Comparing the proposed method with some standard string similarity measures 98
5.4.3 Standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2 100
5.5 Discussion and Conclusion on String Similarity of Domain-Specific Short Text Segments 111

6 Discussion of a Real-World Application in Nutrition Science 113
6.1 Discussion on DSC 113
6.2 Discussion on drNER 118
6.3 Discussion on StandFood 119
6.4 Conclusion 122

7 Conclusions and Further Work 123

Appendix A Two-Sample Kolmogorov-Smirnov Test 127

Appendix B Data Used for Statistical Experiments 129

Appendix C Rules for post-hoc chunkings 131

References 133

Bibliography 149

Biography 151

List of Figures

Fig 3.1: Probability density functions. . . . 21
Fig 3.2: Boxplot of the data for the countries C_2, C_4, and C_{15}. . . . 31
Fig 3.3: Cumulative distributions (step functions) and average values (vertical lines) for f_3 of C_2, C_4, and C_{15}. . . . 33
Fig 3.4: Cumulative distributions (step functions) and average values (vertical lines) for f_{19} of C_2, C_4, and C_{15}. . . . 34
Fig 3.5: Cumulative distributions (step functions) and average values (vertical lines) for f_{21} of C_2, C_4, and C_{15}. . . . 35
Fig 3.6: Cumulative distributions (step functions) and average values (vertical lines) for f_{22} of C_2, C_4, and C_{15}. . . . 35
Fig 3.7: Boxplot of the data for the countries C_8, C_2, and C_{13}. . . . 37
Fig 3.8: Boxplot of the data for the countries C_1, C_3, and C_6. . . . 38
Fig 3.9: Boxplot of the data for the countries C_3, C_1, C_{15}, and C_8. . . . 41
Fig 3.10: Boxplot of the data for the countries C_1, C_{11}, C_2, and C_9. . . . 42
Fig 3.11: Boxplot of the data for the countries C_1, C_6, C_4, and C_3. . . . 42
Fig 3.12: Cumulative distributions for f_{17}. . . . 50
Fig 3.13: Kernel density estimation for the obtained p-values for each pairwise comparison. . . . 53
Fig 3.14: Visualization of post-hoc tests used for C_2, C_4, and C_{15}. . . . 55
Fig 3.15: Visualization of post-hoc tests used for C_8, C_2, and C_{13}. . . . 55
Fig 3.16: Visualization of post-hoc tests used for C_1, C_3, and C_6. . . . 56
Fig 3.17: Number of hypotheses rejected between C_1, C_3, and C_6. . . . 57
Fig 3.18: Average p-value between C_1, C_3, and C_6. . . . 57
Fig 3.19: Number of hypotheses rejected between C_{11}, C_2, and C_{13}. . . . 58
Fig 3.20: Average p-value between C_{11}, C_2, and C_{13}. . . . 58
Fig 3.21: Number of hypotheses rejected between C_1, C_3, and C_6. . . . 59
Fig 3.22: Average p-value between C_1, C_3, and C_6. . . . 60
Fig 3.23: Cumulative distributions (the step functions) and average values (the horizontal lines) for different problems of C_9, C_{14}, and C_1. . . . 61
Fig 3.24: Probability density functions for f_{18} for the countries C_9, C_{14}, and C_1. 62
Fig 3.25: Power analysis for the combination C_6, C_{14}, and C_2. . . . 63

Fig 4.1: DrNER workflow diagram. . . . 71
Fig 4.2: Karnaugh map of the Boolean function for the first post-hoc chunking. 74
Fig 4.3: Graphic representation of the recommendation Φ_1. . . . 82
Fig 4.4: Parse tree of Φ_1. . . . 82
Fig 4.5: Distribution of documents per institution. . . . 85
Fig 4.6: Distribution of number of sentences per documents from scientifically validated web sites. . . . 86
Fig 4.7: Distribution of number of sentences per documents that are abstracts from scientific publications. . . . 86

Fig 5.1: StandFood classification part flowchart. 103
Fig 5.2: StandFood description part flowchart. 104
Fig 5.3: StandFood post-processing part. 105

Fig 6.1: Histograms of FFQ answers of each group for f_3, f_4, and f_8. 116
Fig 6.2: E-learning tool: Advanced Statistics in Natural Sciences and Technologies. 117
Fig 6.3: E-learning tool: Deep Statistical Comparison. 118
Fig 6.4: Word cloud of recipes ingredients. 120
Fig 6.5: Word cloud of ingredients for which we found analyses in the EuroFIR database. 121
Fig 6.6: Word clouds of ingredients for which we found analyses in databases of European countries. 122

List of Tables

Table 2.1: Classification of different statistical tests. 13

Table 3.1: Statistical comparisons of three countries. 30
Table 3.2: Averages for multiple problem analysis obtained by the common approach for the countries: C_2, C_4, and C_{15}. 31
Table 3.3: Two-sample *Kolmogorov-Smirnov test* between the pairs of countries. . . 32
Table 3.4: Ranks obtained by the DSC ranking scheme for the countries C_2, C_4, and C_{15}. 32
Table 3.5: Ranks obtained by the *Friedman test* using the common approach with averages for the countries C_2, C_4, and C_{15}. 33
Table 3.6: Ranks obtained by the *Friedman test* using the common approach with medians for the countries C_2, C_4, and C_{15}. 36
Table 3.7: Ranks for the countries C_8, C_2, and C_{13}. 38
Table 3.8: Ranks for the countries C_1, C_3, and C_6. 39
Table 3.9: Statistical comparisons of three countries using the *Friedman aligned-ranks test*. 39
Table 3.10: Statistical comparisons of four countries. 40
Table 3.11: Ranks obtained by the *Friedman test* using the common approach with averages for 10 countries. 43
Table 3.12: Ranks obtained by the DSC ranking scheme for 10 countries. 44
Table 3.13: Multiple comparisons with a control country (C_5) by using the average ranks obtained by the *Friedman test* with averages. 44
Table 3.14: Multiple comparisons with a control country (C_5) by using the average ranks obtained by the DSC ranking scheme. 45
Table 3.15: Ranks obtained by the DSC ranking scheme without FWER control for 10 countries. 46
Table 3.16: Multiple comparisons with a control country (C_5) by using multiple *Wilcoxon* tests with the DSC ranking scheme. 46
Table 3.17: Statistical comparison of three countries using DSC sequential ranking scheme. 47
Table 3.18: Ranks for the countries C_{11}, C_8, and C_{15}, using the first version of the *DSC* ranking scheme. 48
Table 3.19: Permutations of independent measurements obtained on the problem f_{17} for the countries C_{11}, C_8, and C_{15}. 51
Table 3.20: Statistical comparison of three countries using DSC. 52
Table 3.21: Ranks for the countries C_9, C_{14}, and C_1. 60

Table 4.1: The Boolean function for the first post-hoc chunking. 73
Table 4.2: USAS categories. 77
Table 4.3: Dictionaries. 77
Table 4.4: Default chunking result for one sentence. 78

Table 4.5: Example of the first post-hoc chunking. 79
Table 4.6: Matrix that contains candidates for each entity. 79
Table 4.7: The first phase of the drNER method for Φ_1. 81
Table 4.8: Information extraction of 13 dietary recommendations 83
Table 4.9: Evaluation results. 87
Table 4.10: Performance measures. 88

Table 5.1: Food matching results. 100
Table 5.2: Classification accuracy for each ML algorithm using 10-fold cross validation. 106
Table 5.3: Precision and recall for each food category using the evaluation set. . . . 107
Table 5.4: Correctly classified instances by the StandFood classification part. 108
Table 5.5: Result from the StandFood description part for ten randomly selected instances. 109
Table 5.6: Precision and recall for each food category after applying the post-processing rules. 110
Table 5.7: Food categories for seven food items after post-processing rules. 110

Table 6.1: Ranks for population groups. 115

List of Algorithms

Algorithm 3.1: Pseudocode of the DSC ranking scheme used to rank data sets on the i-th problem. 25
Algorithm 4.1: Pseudocode of drNER. 72

Abbreviations

AD	...	Anderson-Darling
AL	...	Active Learning
AnEM	...	Anatomical Entity Mentions corpus
ANOVA	...	Analysis of Variance
API	...	Application Programming Interface
ASO	...	Alternating Structure Optimization
BANNER	...	Named-Entity Recognition for Biomedical Text
BB3	...	Bacteria Biotopes 3
BioC	...	Simple format to share text data and annotations
B-NP	...	Beginning of Noun Phrase
CD	...	Cardinal Number
CHEMDNER	...	Chemical and Drug Name Recognition
CRAFT	...	Colorado Richly Annotated Full Text corpus
CRF	...	Conditional Random Field
CI	...	Computational Intelligence
DSC	...	Deep Statistical Comparison
DNN	...	Deep Neural Network
DRVs	...	Dietary Reference Values
EFSA	...	European Food Safety Authority
EuroFIR	...	European Food Information Resource
FAO	...	Food and Agriculture Organization
FBDGs	...	Food-Based Dietary Guidelines
FCDB	...	Food Composition Database
FFQ	...	Food Frequency Questionnaire
FoodEx2	...	Standardized system for classifying and describing food
FWER	...	Family-Wise Error
GE4	...	Genia 4
GIFT	...	Global Individual Food consumption data Tool
GM	...	Gene Mention
IE	...	Information Extraction
I-NP	...	Inside of Noun Phrase
JJ	...	Adjective
JJR	...	Adjective, comparative
JJS	...	Adjective, superlative
KS	...	Kolmogorov-Smirnov
MaxEnt	...	Maximum Entropy
ML	...	Machine Learning
NCBI	...	National Center for Biotechnology Information
NER	...	Named-Entity Recognition
NLP	...	Natural Language Processing
NN	...	Noun, singular or mass

NNET	...	Neural Networks
NNS	...	Noun, plural
NNP	...	Proper noun, singular
NNPS	...	Proper noun, plural
POS	...	Part-Of-Speech
PP	...	Preposition Phrase
RDF	...	Resource Description Framework
RF	...	Random Forest
SeeDev	...	Genetic and molecular mechanisms involved in plant Seed Development
SLDA	...	Scaled Linear Discriminant Analysis
SVM	...	Support Vector Machine
TREE	...	Classification tree
VB	...	Verb, base form
VBD	...	Verb, past tense
VBG	...	Verb, gerund or present principle
VBN	...	Verb, past participle
VBZ	...	Verb, 3rd person singular present
WHO	...	World Health Organization

Chapter 1

Introduction

The thesis proposes a new methodology that is a synergy of statistical data analysis and natural language processing to explore a given domain, which in our case is the nutrition science. In this chapter, we provide the motivation and the problem definition, hypotheses, and the scientific contributions. In addition, we present the structure of the thesis.

1.1 Problem Definition

Nutrition science that includes clinical nutrition, food and nutrition management, public health nutrition, personalized nutrition, etc., has its foundation in the biological, chemical, and medical sciences, with a focus on nutrient/non-nutrient function and metabolism. One of the main objectives of the nutrition science is to establish food-based dietary guidelines (FBDGs) to achieve optimum health, and treatment or prevention of disease conditions as well as food production and safety [1]. FBDGs are simple advices on healthy eating, aimed at the general public, which give an indication as to what a person (an individual member of the general public) should be eating in terms of foods, and provide a basic framework to apply when making healthy dietary choices and planning meals. Even though FBDGs are simple messages, we must be aware that they are based on complex scientific facts, which include dietary reference values (DRVs).

DRVs are nutrient recommendations and quantitative reference values for nutrition intakes, such as population reference intake, the average requirement, adequate intake level, and the lower threshold intake [2]. Authorities (e.g., European Food Safety Agency (EFSA) [3]) continuously identify and review the latest scientific studies, including reports of national and international authorities, for possible health effects of specific nutrients. For example, if the focus is on dietary fiber, health effects of dietary fiber are identified by reviewing scientific studies. Then, evidence of relationships between the intake of a nutrient and health outcome is evaluated. Finally, when nutrient-health relationships are established, the authority provides scientific advice that can be used by policy makers. In practice, this means that a daily intake of 25 g of a dietary fiber is set as a DRV because it is adequate for adults, while consuming more than 25 g of dietary fiber per day may reduce the risk of coronary heart disease and type 2 diabetes and may improve weight maintenance [3].

Most countries have established their own national DRVs that consider, beside international recommendations and guidelines, also local conditions and national/ethnical eating culture and habits, and are reviewed and updated from time to time. A comprehensive review of micronutrient recommendations in Europe, collected within the EU-funded project EURRECA (EURopean micronutrient RECommendations Aligned) [4], was published several years ago [5]. In 2015, the non-profit association EuroFIR [6] updated EURRECA

micronutrient recommendations, enriched them with reference values for other nutrients, and developed a web service for accessing DRVs through the Quisper server platform [7], aimed at collecting scientifically-validated food-related data and knowledge services for dietary advising. However, these recommendations may already be outdated as the nutrition science is a fast growing research area. Beside DRVs for the public, disease-specific DRVs also exist aimed at increasing the awareness of clinicians and persons with chronic diseases about beneficial nutrition therapies. Recently, personalized DRVs have become relevant as they consider genetic predisposition to chronic disease and phenotype information on anthropometry, physical activity, clinical parameters, and biochemical markers of nutrition status, and give strategies to dramatically reduce the risk of chronic-disease. The EU-funded Food4Me project [8] performed a pan-European study of over 1,500 participants, which showed that personalized advice is more effective at improving dietary behavior compared to conventional, population-based FBDGs. In personalized nutrition, knowledge extraction (DRVs) is even more complex because several aspects need to be taken into account.

Once FBDGs are established considering DRVs, food intake needs to be translated into nutrient intake in order to verify its compliance with the guidelines and recommendations. To translate food intake into nutrient intake and reverse, the information about food composition and food consumption data needs to be linked. Food composition databases (FCDBs) provide detailed information on nutrition composition of foods, usually from a particular domain [9]. They usually contain information about a huge number of components, including: energy, macronutrients (e.g., protein, carbohydrate, fat) and their components (e.g., sugars, fatty acids), minerals (e.g., calcium, iron, sodium), and vitamins. Food composition data is essential to many fields, especially in public health and nutrition, but also it has some limitations: 1) variability in the composition of foods between countries; 2) incomplete coverage of foods and nutrients leading to missing values; 3) data validity (some values are out of date). Food consumption data consists of subjects that record all food and beverages consumed over some period of time [10]. Different dietary assessment methods can be used to store information about food consumption data including diet records, 24-hour recall, food frequency questionnaire (FFQ), and innovative dietary assessment methods [11]. Accurately assessing dietary intake is an important element because diet is a major lifestyle-related risk factor of various chronic diseases. The translation has been performed manually, which is a time-consuming task.

In public health, as well in clinical practice, dietary recommendations should rely on evidence-based principles, considering scientific knowledge, expert consensus, and clinical experience that are proved using statistical analyses. However, many scientists have problems and difficulties in making statistical analysis of their data, which is often crucial in interpreting their results. The problem appears because each statistical method has some conditions (assumptions) about the data that must be satisfied in order to apply it. Usually, these initial conditions are not checked and researchers simply apply a statistical method, in most cases taken from a similar published work, which is unsuited to their data. As a result, their conclusions can be incorrect. A good tutorial on how to perform statistical analysis is provided by Demšar in 2006 [12], where the main focus was statistical comparisons of classifiers over multiple data sets. This work was followed by several works done by García and Herera in 2008, 2009, 2010, and 2011 [13]–[16], where the focus was on statistical comparisons in Computational Intelligence (CI). These works are generally applicable in any domain. Also, one scenario working with nutrition experimental data is when one parameter is measured by different methods and needs to be compared over multiple samples and each sample is measured several times. A statistical analysis, followed by some state-of-the-art literature, uses the average value of the parameter obtained by each

method on each sample. However, averages are sensitive to outliers and consequently medians are sometimes used. In classical inferential statistics using the state-of-the-art methods with either averages or medians, the results can be affected by the ranking scheme that is used by some standard statistical tests. According to this, statistical analyses in nutrition science should be carefully made because the reported knowledge is the basis for FDBGs.

Existing resources of DRVs and FBDGs consist of structured and unstructured information that needs to be updated with the new scientific research. For example, searching the PubMed library for the keyword "nutrition", more than 20,000 relevant scientific papers published in 2017 can be found and it is impossible to read all of them in order to extract the new knowledge related to dietary recommendations. To update the information about the dietary recommendations, the information presented in recently published literature and scientifically validated public health web sites needs to be exploited. These resources have a lack of coded data (e.g., unique identifiers from ontologies), but on the other hand have a lot of unstructured text data that needs to be analyzed to correctly interpret the dietary information. The amount of information presented as unstructured text is massive and is quickly increasing, so computer-based methods for systematic knowledge identification, extraction and exploration are welcome to support human experts decision-making about the appropriate nutrition care for specific disease states or conditions in typical settings. Automatic identification and classification of words or phrases that describe important concepts (entities) can be done by a process known as Named Entity Recognition (NER) [17]–[19]. For example, in the last 10 years, different shared workshops were organized for extracting information from biomedical literature. NER methods were presented for gene entities recognition, protein entities recognition, extracting relationships between genes and proteins, phenotype information, disease-phenotype relationships, chemical entities, disease-drug relationships, etc., but to the best of our knowledge there is no research that has been focused on the extraction of dietary information concepts and on the relations that exist between them. Several studies are conducted in the dietary domain, but with different goals [20]. Also, most of the existing NER methods are developed using Machine Learning (ML) [21]–[23] algorithms based on an annotated corpus in the domain of interest, which is provided by human experts form the domain and it requires time and effort to produce it. To the best of our knowledge, in the domain of dietary recommendations, an annotated corpus provided by the experts is still missing.

The thesis addresses three problems. The first one is how to obtain more robust results from statistical analysis of experimental data, especially in the case when one parameter is measured by different methods and needs to be compared over multiple samples, and each sample is measured several times. The second one is focused on the extraction of the relevant information about dietary domain that is scientifically published and a result of statistical analysis. The third one is focused on linking the extracted food intake to nutrient intake or reverse.

For the first problem, the novelty of the method will rely on a ranking scheme that is based on the whole distribution, instead of using only one statistic to describe the distribution of the data, such as average or median. The method is generally applicable in any domain including: chemical analyses, biological analyses, physical analyses, economy, evolutionary optimization, machine learning, etc.

The novelty of the second method will rely on a NER that is not based on an annotated corpus. It is completely Natural Language Processing (NLP) [24] method that uses rules, which are not based on characteristics of the entities of interest. It can be easily updated for some other domains for which an annotated corpus does not exist.

The novelty of the last method will rely on string similarity measures of short text segments. By combining the morphological information together with the probability theory

promising results can be achieved.

These methods will significantly contribute to the improvement of the quality of food- and nutrition-related data and knowledge and also can be used for other domains.

1.2 Hypotheses

The hypotheses of the thesis relate to methods for exploring a given domain, or in our case food- and nutrition-related data and knowledge. The methods focus on obtaining more robust results from statistical analysis, automatically extracting the results reported in scientific publications, and normalizing the extracted knowledge. The first hypothesis is related to statistics, the second hypothesis is related to natural language processing and is split into two parts, the first part is related to information extraction and the second part is related to text normalization, and third hypothesis is related to nutrition science:

- **Hypothesis 1:** In classical inferential statistics, the results of the state-of-the-art methods for statistical comparisons when one parameter is measured by different methods and needs to be compared over multiple samples, and each sample is measured several times, can be affected by the presence of outliers and a small epsilon-neighborhood that exists between data values that are compared. Our first hypothesis is that, in order to obtain more robust results, the whole distribution of the data should be taken into account in the ranking process instead of using a simple statistic such as average or median, which are commonly-used to describe the data.

- **Hypothesis 2A:** To extract information from text, different NER methods can be used. Most of them are corpus-based NERs while the other are rule-based NERs where rules are created using the characteristics of the entities of interest. One part of our second hypothesis is that, to extract information from an untapped domain, a rule-based NER method can be used without an annotated corpus, where the rules will not be associated to the characteristics of the entities of interest.

- **Hypothesis 2B:** String similarity measures are the basis of text normalization methods. The other part of our second hypothesis is that by probability modeling of the domain using the morphological information, string similarity measures of domain specific short text segments can be improved.

- **Hypothesis 3:** With the revolutionary developments in life science technologies, characterized by high throughput, high efficiency, and rapid computation, as well as access to bioinformatics databases, computer-based methods for systematic knowledge identification, extraction and exploration are welcome to improve the nutrition science analyses. Our third hypothesis is that all of the proposed methods can improve the quality of some analyses performed in nutrition science.

1.3 Scientific Contributions

The work presented in this thesis leads to several contributions to the areas of statistics, natural language processing, and food and nutrition. A complete list of publications related to this thesis is available in the Bibliography section. Each contribution is related to one of the hypotheses outlined above:

- *An approach to statistical comparison of experimental data that is more robust on outliers and small differences that can exist between data values using the whole distribution of the data instead of using only one statistic to describe the distribution,*

such as average or median. This work has been presented at conferences [25]–[28] and a journal article [29].

Working with experimental data a crucial task is the collection, analysis, interpretation and presentation of data. For this reason, researchers should perform a statistical analysis to correctly interpret the results of the study. However, many scientists working not only in the field of food and nutrition but also in other fields have problems and difficulties in making a statistical analysis of their data. This happens considering the fact that they do not check the conditions that need to be satisfied in order to use some statistical method, but they select a statistical method from a similar study. A common scenario is when a parameter is measured by different methods over multiple samples, and each sample is measured several times. In this case, state-of-the-art literature suggests using an average or median of multiple measurements as a unique value for each method on each sample. However, using averages and medians could affect the results of the study. For this reason, we propose a novel approach for making a statistical comparison of experimental data that is more robust on outliers and small differences that can exist between data values. The main contribution of the approach is at the ranking scheme, which is based on the whole distribution, instead of using only one statistic to describe the distribution, such as average or median. The experimental results show that our approach gives more robust results compared to the commonly-used approach in cases when the results are affected by outliers or by a misleading ranking scheme. The approach is generally applicable in any domain including: chemical analyses, biological analyses, physical analyses, economy, evolutionary optimization, machine learning, etc.

- *A rule-based named-entity recognition method for information extraction, which does not use an annotated corpus and the rules are not associated with the characteristics of the entities of interest* and its application to the dietary domain, especially evidence-based dietary recommendations. This work has been presented at conferences [30], [31] and published in a journal article [32].

  The results of a study are usually presented in scientific publications. Nowadays to follow the new knowledge in a particular domain, which is quickly increasing with recently published papers, computer-based methods are extremely welcome for knowledge identification, extraction, and exploration. In the last 10 years, the NLP community organizes shared workshops where the focus is on information extraction for a particular domain from biomedical literature. Most of the existing NER methods are developed using ML algorithms based on an annotated corpus in the domain of interest, which is provided by human experts from the domain and it requires time and effort to produce it. We develop a rule-based named-entity recognition method for information extraction that does not use an annotated corpus and the rules are not associated with the characteristics of the entities of interest. It is developed for the dietary domain, which is still an untapped domain. Nowadays, the need of exploring the dietary domain is really important to follow the new knowledge that comes in the nutrition science. For example, in mid December 2014, the JPI HDHL European Nutritional Phenotype Assessment Data Sharing Initiative (ENPADASI) officially started, where the main objective is to deliver an open access research infrastructure that will contain data from a wide variety of nutrition studies, ranging from interventions to epidemiological studies including a multitude of phenotypic outcomes that will facilitate the combined analysis in the future. In autumn 2015, the European Union's Horizon 2020 project, RICHFIELDS, started, which aims to design a consumer-data platform to collect and connect, compare and

share information about our food behaviors, to revolutionize research on every-day choices made across Europe. The purpose of these infrastructures is to harmonize different food- and nutrition-related data, which will be further available for analysis. However, some data that needs to be collected by these infrastructures is presented in unstructured form, so first it should be extracted and then analyzed.

- *A methodology for string similarity of domain-specific short text segments*, its implementation and evaluation on the task of food matching, or matching food consumption and food composition data that can be described and classified in different ways. These have been presented in conference proceedings [33]–[37] and published in a journal article [38].

  The information of interest may be mentioned in text in multiple ways, using phrases with a variety of structures. The alternative ways in which the same concept may be mentioned creates a problem to the automatic integration of information. For this reason, there is a need of an automatic linking of concept mentions to known concept from a domain-specific terminological resource. Normalization methods are used for these problems, which are based on string similarity measures. For this purpose, we develop a methodology for a string similarity of domain-specific short text segments that combine morphological information with the probability theory. Compared to other text similarity measures that can fail in string similarity between two very short text segments, the proposed methodology gives promising results. We consider a real problem of food matching, which is one of the most interested topics for 2017 in food science. The problem that appears here is that the name of the same food can be different regarding different ways of how people express themselves. To allow automatic transformation from food intake to nutrient intake the same food needs to be linked to a concept that exists in the food information systems. By using the proposed methodology promising results can be achieved.

- *Computer-based methods for systematic knowledge identification, extraction and exploration*, related to some fundamental tasks that are recently under research in nutrition science, such as statistical data analysis, information extraction, and food matching (food data harmonization). These have been presented in conference proceedings [34]–[37], in a book chapter [39] and published in a journal article [38].

  Statistics has become an increasingly important tool to quantitatively analyze information in nutrition science. Nowadays, almost each paper published in nutrition science journal contains a section related to statistical data analysis, which is a basis for the discussion and the results of the study. Many researchers have problems and difficulties performing a statistical analysis of their data and this happens according to the questions related to the selection of an appropriate statistical method and the statistic that will be used as measure. To become familiar with statistical comparison, we provide a short tutorial on how to perform a statistical analysis of food data, presented as a book chapter in the "Science within Food: Up-to-date Advances on Research and Educational Ideas" [39]. In addition, we developed an e-learning tool, which is available at `http://ws.ijs.si/statTool/` and it not only reduces the time needed to perform a statistical analysis but importantly, it can help in interpreting results and increase awareness of using an inappropriate statistical method. The e-learning tool was also presented at the ISO-FOOD Spring School on Food Authentication by Isotope and Elemental Fingerprinting.

  Over the past few years, information extraction in nutrition science, becomes a topic that is going under research in several European projects: "Era-Chair ISO-FOOD",

"ENPADASI", and "RICHFIELDS", where the main goal is development of research infrastructure or knowledge repository in order to collect and connect information in nutrition science. For this reason, as a part of "Era-Chair ISO-FOOD" and "RICHFIELDS", we developed drNER, which can be used for information extraction of evidence-based dietary recommendations. DrNER was also presented at the 2017 Annual Meeting of the International Society of Behavioral Nutrition and Physical Activity, as a part of the Symposium entitled "Man or machine? How far are we in the field of smart devices for dietary data collection & analysis?" [34].

Another topic that is going under nutrition science research, especially in 2016 and 2017, is the coding of food consumption data and food composition data in the same food system, which will reduce errors and time needed for food matching. The food matching is an important task that need to be done when more data sets need to be combined in order to answer questions about the source of nutrients in people's diet, the level of consumption among high consumers of different foods, and the food eating choices. In November 2017, the Food and Agriculture Organization of the United Nations, FAO, and the World Health Organization, WHO, lunched a FAO/WHO Global Individual Food consumption data Tool (FAO/ WHO GIFT, `http://www.fao.org/gift-individual-food-consumption/en/`), which provides up-to-date inventory of individual quantitative food consumption surveys conducted in low- and middle-income countries, where all food consumption data are coded in FoodEx2, which is a system for classifying and describing foods. The platform also provides an opportunity to make other food consumptions surveys visible, which also need to be FoodEx2 coded. Finding the FoodEx2 codes is a time-consuming task because usually it is manually made and also requires a good knowledge of the FoodEx2 system. However, we previously developed a system, known as StandFood, which is a semi-automatic system for standardization of foods according to FoodEx2. It significantly reduces the time needed to find the missing FoodEx2 data. StandFood was used for matching foods from EuroFIR food composition databases to FoodEx2 [36], [38]. It was also presented at the third International Conference of Metrology Promoting Standardization and Harmonization in Food and Nutrition [35]. It is the first semi-automatic published system for standardization of foods according to FoodEx2, while all other approaches are based on manual food matching. A part of StandFood was also used to map food composition data from EuroFIR and USDA food composition databases in order to help the process of borrowing missing values that exist in food composition databases [37].

Each of the proposed methods is developed to solve different task, but further, the results obtained from them can be combined for some more advanced analysis, such as automatic dietary assessment or to plan and implement national food and nutrition programmes and interventions.

1.4 Thesis Structure

In the Introduction, we have provided an overview of the basic concepts in the food and nutrition domain, followed by the limitations of existing approaches that can be used to explore this domain, our research hypotheses to overcome these limitations, and the scientific contribution for each research hypothesis, separately.

Chapter 2 presents the background that is crucial for understanding the methods that are presented in this thesis. The focus is on two different areas. First, we focus on statistics which is a crucial task for data analysis of experimental data, which is needed before the results of a study are published. From statistics, we review some basic concepts

and the idea of hypothesis testing. Second, we give an overview of methods used for information extraction of the results of studies published in scientific papers. We give a brief overview of natural language processing, machine learning, named-entity recognition, and text normalization methods.

Chapter 3 is related to the first research hypothesis and presents a novel approach to statistical comparison of experimental data that is more robust on outliers and small differences that can exist between data values. It is focused on how to obtain more correct results from experimental data that should be published.

Chapter 4 is related to the first part of the second research hypothesis and presents a rule-based named-entity recognition method for information extraction that does not use an annotated corpus and the rules are not associated with the entities characteristics. Its evaluation is performed on an untapped domain, the dietary domain, especially evidence-based dietary recommendations.

Chapter 5 is related to the second part of the second research hypothesis and presents a methodology for a string similarity of domain specific short text segments. We illustrate the advantages of the novel proposed methodology on the food matching task, or linking food consumption and food composition data.

Chapter 6 presents a discussion of real-world application in state-of-art nutrition science in order to present the benefits of the proposed methods for nutrition science, which is also related to our last research hypothesis.

Chapter 7 presents the overall discussion of the proposed methods and the scientific contributions achieved. Finally, we conclude the thesis and give some directions for future work.

Chapter 2

Background

In this chapter, we introduce the background that is crucial for understanding the methods that are presented in this thesis. It consists of two parts. The first one gives an overview of the basic terms used in statistics with a special focus on statistical comparison. The second one is related to information extraction, which includes presenting the basic ideas of natural language processing (NLP) and machine learning (ML), followed by the explanation of the named-entity recognition task and text normalization task. The related work for each method is further elaborated in the chapters where the method is explained.

2.1 Statistical Comparison

Statistics can be interpreted in various ways. According to the most used definition, statistics is a branch of mathematics that is related to the collection, analysis, interpretation or explanation, and presentation of data [40]. In general, it is used for analyzing and summarizing a given data set, which can consist of experimental data or data from real-life studies. The basic idea behind it is that by analyzing a sample of a larger population, interpretations about the population can be developed using the outcomes from the sample. It is used particularly in any domain that involves data, such as nutrition science, psychology, business, physics, chemistry, biology, computer science, government, manufacturing, etc.

Classical statistics distinguishes between two types of statistical analyses: *descriptive statistics* and *inferential statistics*. To explain them, we introduce some basic terms used in statistics, followed by a special focus on *hypothesis testing*.

2.1.1 Basic terms

A *population* is a set of similar items or events, which are of interest for some question or experiment, while a *representative data sample* is a set of data collected and/or selected from a statistical population by a defined procedure [40]. To see the difference between them, let us consider that a researcher is interested in how many people drink milk with breakfast in Slovenia. In this example, the population contains each person who lives in Slovenia. It would be unrealistic for us to ask each individual about his or her milk drinking habits, so instead we need a representative sample of people. The information obtained from the representative data sample allows a researcher to develop hypotheses about the larger population. In this example, let us consider that the researcher is a vegan and randomly selects a sample in which many of the participants are his or her friends many of whom are vegans. In this case, the number of people who drink milk with their breakfast will be lower when in reality the number is larger. This is an example of sample selection bias. In order to have a representative data sample, the researcher needs to be

unbiased in their selection i.e. there are no outside factors influencing the sample selection. One way to obtain a representative sample is to use random sampling, which involves a random selection of the instances from the target population and minimizes the bias. If the sample is not representative, the result is known as sampling error.

Once a representative data sample was selected, it becomes important to know what type of data has been collected. There are two types of data: *qualitative* (categorical) and *quantitative* (numeric). Each type is split into two sub-types: *ordinal* and *nominal* for qualitative data, and *discrete* and *continuous* for quantitative data. *Ordinal* data is categorical, where the values have natural, ordered categories and the distances between the categories are not known. For example, a question from a food questionnaire might ask: "How often do you eat fish during the week?". Possible answers include "Never", "One to three times per week", "Four to six times per week", or "Every day". In this case, the answers are examples of ordinal data. *Nominal* data is also a categorical data, but in this case there is no natural order between the categories. Examples include eye colour, gender, and the region where one lives. In contrast to qualitative data, quantitative data is numerical. *Discrete* data can only take on a finite number of numeric data, while *continuous* data can take on an infinite number of possible values. Let us assume that a researcher is interested in the weight of participants in the data sample. The weight of each participant is stored as a continuous variable. However, he/she can split the participants into four groups according to their weight: (1) less than 50 kg, (2) 50 kg - 65 kg, (3) 66 kg - 80 kg, and (4) greater than 80 kg. In this case, weight is a discrete variable because it can take only one value from the four weight groups. This is an example of discretization, where a continuous variable is transformed into a discrete variable.

After the data is collected, the next step is to apply statistics to the representative data sample. As we said before, there are two different statistical analyses that we can apply: descriptive statistics and inferential statistics [41]. *Descriptive* statistics summarizes data from a sample using measures, while *inferential* statistics draws conclusions from the data set subject to random variation.

Descriptive statistics includes a *distribution* of a single variable, measures of *central tendency*, which include the *mean*, *median*, and *mode*, and measures of *variability*, which include *standard deviation*, *variance*, and the *minimum* and *maximum value*. The *distribution* is a summary of the frequency of individual values or a range of values for a variable. It represents every value of the variable and the number of how many times the value appears in the data sample. The central tendency of a distribution is an estimate of the "centre" of a distribution. The *mean* (average) is the most commonly used measure for central tendency and in order to compute it we sum all the values in our data sample and then divide the sum by the sample size (number of values in the data sample). The *median* is the value that separates the higher half of a data sample from the lower half, or it can be assumed as the "middle" value of an ordered data sample. The *mode* is the value from a data sample that appears with the highest frequency. The variability of a distribution refers to the spread of the data values around the central tendency. The *minimum* and the *maximum* value are the minimum and the maximum value that appear in the data sample. The *standard deviation* is used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data values are close to the mean, while a high standard deviation indicates that the data values are spread out over a wider range of values. The *variance* is the square of the standard deviation.

Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected to represent the population perfectly. The methods of inferential statistics include estimation of distribution parameters and the testing of statistical hypotheses.

When we talk about statistical analysis, it is also important to distinguish between *univariate*, *bivariate* and *multivariate* statistical analysis. Univariate analysis works when the data sample only has one variable. It does not deal with relationships and the major purpose is to describe the data. Bivariate analysis involves analysing two variables in order to find a relationship between them. Bivariate analysis is a special case of multivariate analysis when multiple relations between multiple variables are analysed. To explain better the difference between them, let us assume that the researcher is interested in testing whether or not consulting with a nutritionist can decrease a person's calorie intake. For this purpose, a researcher works with participants who consult with a nutritionist each day after work. The participants form two groups: one group that has daily consultations and a control group who do not receive any nutritional advice. After a week, let us assume that the researcher discovered that those who consult with a nutritionist have decreased their calorie intake for 40% over the control group. This is an example of univariate inferential statistics because it analyzed the relationship between one independent variable (consulting with a nutritionist) with a single dependent variable (calorie intake). If a researcher is interested in the relationship between the times a person consulted a nutritionist and calorie intake over the week, then he/she needs to perform a bivariate analysis. If the researcher is interested in adding another variable i.e. how many times a person visited a gym, to determine the effectiveness of visiting a nutritionist and exercise, on two groups of people (20-40 yrs., and 41-60 yrs.), then he/she needs to perform multivariate analysis to find the relationships between all of the variables.

2.1.2 Hypothesis testing

One of the most commonly used approaches for testing relationships between two or more data samples is to use hypothesis testing [42], or it is a procedure in which sample(s) data is employed to evaluate a hypothesis. Hypothesis testing, also called significance testing, is one method that can be used to test a hypothesis about a parameter in a population, using data measured in a data sample, or about the relationship between two or more populations, using data measured in data samples. The method starts by defining two hypotheses, the *null hypothesis* H_0 and the *alternative hypothesis* H_A. The null hypothesis is a statement that there is no difference or no effect and the alternative hypothesis is a statement that directly contradicts the null hypothesis by indicating the presence of a difference or an effect. This step is crucial, because mis-stating the hypotheses will disrupt the rest of the process. We can define the hypotheses about a single population or about the relationship between two or more populations e.g., between either the means or variances of a single, two or multiple distributions or even about the whole distribution. The next step is to select an appropriate *test statistic* T, which is a measurable function of a random sample that allows researchers to determine the likelihood of obtaining the outcomes if the null hypothesis is true. The level of significance α, also called the *significance level*, which is the probability threshold below which the null hypothesis will be rejected, also needs to be selected. The last step is to make a decision either to reject the null hypothesis in favor of the alternative or not to reject it. This decision can be made using two different approaches. In the first, all the possible values of the test statistic for which the null hypothesis is rejected, also called the *critical region*, are calculated using the distribution of the test statistic and the probability of the critical region, which is the level of significance α. Then the observed value of the test statistic T_obs is calculated according to the observations from the data sample. If the observed value of the test statistic lies within this critical region, the null hypothesis is rejected, and if not, it is not rejected. In the second approach, instead of defining the critical region, a *p-value* i.e, the probability of obtaining the sample outcome, given that the null hypothesis is true, is calculated. The null hypothesis is rejected, if

the p-value is less than the significance level (typically 0.05 and 0.1), and if not, it is not rejected.

In reality, the null hypothesis can be either true or false, and the result of a statistical test can be that the null hypothesis is either rejected or it is not rejected. When performing hypothesis testing, two types of errors can occur: *type I* and *type II*. A *Type I* error occurs if we reject the null hypothesis when it is true ($\alpha = P(Type\ I\ Error)$). A *Type II* error occurs if we fail to reject the null hypothesis when the alternative hypothesis is true ($\beta = P(Type\ II\ Error)$). The probability of a type I error is the level of significance (α), so before the study we usually assign it a small value (e.g., 0.05, 0.01) because researchers do not want to have a type I error. The probability of type II error is marked as β and it is related to the *power* of a statistical test. The power is the probability that it will reject a false null hypothesis, or power $= 1 - \beta$ [43].

Power analysis is an important aspect in experimental design [44]. It allows researchers to determine the sample size required to detect an effect of a given size with a given degree of confidence. It allows researchers to find the probability of detecting an effect of a given size with a given significance level, under sample size constraints. If the probability is very low, then researchers can change the sample size of the experiment.

Two types of statistical tests exist: *parametric* and *nonparametric*. In order to select an appropriate statistical test, and to choose between a parametric and a nonparametric test [14], the first step is to check the assumptions of the parametric tests, also called the required conditions for safe use of parametric tests. These conditions include checking for *independence*, *normality of the data*, and *homoscedasticity of the variances*. Statistically, two events are *independent* if the occurrence of one does not influence the probability of the other occurring. *Normality* indicates that the data is normally distributed, which we can check by using statistical tests such as the *Kolmogorov-Smirnov* [45], *Anderson-Darling* [46], *Shapiro-Wilk* [47], and *D'Agostino-Pearson* test [48]. The difference between these statistical tests and the recommendations which one is more appropriate for a typical scenario are given in [49]. The result from a statistical test can be graphically proven using *histograms* or *Q-Q plots (quantile-quantile)*. In probability, quantiles are cut points dividing the range of a probability distribution into intervals with equal probabilities. The *homoscedasticity of variances* indicates the hypothesis of equality of variances (homogeneity of variances). The *Levene's* test [50] can be used to check the homoscedasticity of the variances. If the data does not satisfy the required conditions for safe use of parametric tests, then tests can result in incorrect conclusions, and it is better to use an analogous nonparametric test. In general, a nonparametric test is less restrictive than a parametric one, but it is also less powerful than a parametric one, when the required conditions for safe use of parametric test are satisfied [14].

In addition to the conditions for the safe use of parametric tests, other parameters that are also related to the selection of an appropriate statistical test are the number of data samples that need to be compared (two or more than two) and if the data samples are *paired* or *unpaired*. *Paired* samples (also called dependent samples) are samples in which natural or matched couplings occur. So in the data sample each data value in one sample is uniquely paired to a data value in the second sample. Examples of paired samples are found in food questionnaires. Let us suppose that researchers are interested if there is a difference between two populations (males and females) according to the food questionnaire. In this case the obtained result of the male population for the first question is paired with the obtained result of the female population for the same question, and so on. The choice between paired and unpaired samples depends on experimental design, and researchers need to be aware of this when designing their experiment.

Table 2.1 presents different statistical tests classified according to the conditions that

Table 2.1: Classification of different statistical tests.

	Two data samples	**More than two data samples**
Parametric	t-test (unpaired)	One-way ANOVA (unpaired)
	Paired t-test (paired)	Repeated-measures ANOVA (paired)
Nonparametric	Mann-Whitney U (unpaired)	Kruskal-Wallis (unpaired)
	Wilcoxon signed rank (paired)	Friedman, Friedman-aligned, Iman-Davenport (paired)

must be met.

The *t-test* is a parametric statistical test that can be used to determine if two unpaired data sets of data are significantly different from each other [51]. If the required conditions for the safe use of the parametric test are not satisfied, an alternative nonparametric version of the t-test is *Mann-Whitney U test* [52].

The *paired t-test* [53] is a parametric statistical test that can be used for pairwise comparison. It is used to test whether the difference between two data sets over multiple problems is non-random. It checks whether the average difference over the multiple problems is significantly different from zero. To use this test, the required conditions for the safe use of the parametric tests need to be satisfied. The *Wilcoxon signed-rank test* [54] is a nonparametric alternative to the *paired t-test*, which ranks the differences between the two data sets for each problem, ignoring the signs, and compares the ranks for positive and the negative differences. When the required conditions for the safe use of the parametric tests are satisfied, the *Wilcoxon signed-rank test* is less powerful than the *paired t-test*, otherwise it must be used over the *paired t-test*.

The *One-way ANOVA* is a parametric test and is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups [55]. An alternative nonparametric version to it is a *Kruskal-Wallis test* [56].

The *repeated-measures ANOVA* [57] is a parametric statistical test that can be used to test the differences between more than two related samples. The null hypothesis is that all the data sets are the same and the observed differences are random. If the null hypothesis is rejected, a post-hoc test can be used to find out which algorithms actually differ. In the case of *ANOVA*, the *Turkey test* and the *Dunnett test* [12] are used as post-hoc tests. The first is for comparing all data sets with each other, and the second is used for comparisons of all data sets with one control data set. The *Friedman test* [14] is a nonparametric alternative to the *repeated-measures ANOVA*. It ranks the data sets for each problem separately and, in case of ties, average ranks are assigned. The *Friedman aligned-rank test* [14] is a modification of the *Friedman test*. In the *Friedman test* the ranking scheme allows for intra-problem comparisons only, since inter-problem comparisons are not meaningful. But in the cases, when the number of data sets for comparison is small, comparability among the problems is desirable and in such instances the *Friedman aligned-ranks test* is preferred. The *Iman-Davenport test* [14] is also a modification of the *Friedman test*, in which the Friedman's statistic is modified to avoid a conservative undesirable effect of the Friedman's statistic.

If the null hypothesis in the *Friedman test*, *Friedman aligned-ranks test*, or *Iman-Davenport test*, is rejected, a post-hoc test needs to be used. There are different post-hoc tests that can be used in this case, such as *Nemenyi test* [58], *Bonferroni-Dunn test* [59], *Holm procedure* [60], *Hochberg procedure* [61] and *Bergmann-Hommel's procedure* [13], [62]. These are some of the best known post-hoc tests, and a lot of literature for each of them exists, so we are not going to explain them in more details.

In a typical scenario in any domain, statistical analysis can be conducted in two ways: a single-problem analysis and a multiple-problem analysis. A single-problem analysis is a scenario where the data derives from multiple independent measurements on one problem. For example, two different methods can be compared for measuring beta-carotene in apples. In this scenario, each method is used on a number of apples, so we have multiple independent measurements for each method that need to be compared, and the analysis can involve paired or unpaired samples. Despite the single-problem scenario, the multiple-problem scenario is when we compare data samples over multiple problems or we have paired samples. An example of multiple-problem scenario will be when the above mentioned two methods will be used for measuring beta-carotene not only in apples, but also in other food products from the same food group (e.g., pears, apricots, etc.). In this case, as in most published papers on this topic, authors use average results for each problem (food product) to produce a sample of results for each data sample. For example, the value of beta-carotene that will be involved for a given method in apples will be an average or median of the multiple measurements obtained. These two scenarios are applicable in any domain and good examples of them are presented in [12]–[15], for comparing machine learning classifiers and behaviour of evolutionary algorithms, respectively.

2.2 Information Extraction

Nowadays, a large amount of textual information is available in digital form in public repositories on Internet (e.g., online news, scientific publications, social media, etc.). The textual information is presented as unstructured data, which refers to the fact that the data does not have a pre-defined data model. Working with textual data is a challenge because of its variability, the same concepts can be mentioned in different ways regarding the fact how people express themselves and in different writing styles. Information Extraction (IE) is a task of automatically extracting information from unstructured data and in most cases concerns processing of human language texts by means of NLP [63], [64]. The idea behind IE is to provide a structured representation of extracted information obtained from analyzing text. The information to be extracted is predefined by users, and consists of predefined concepts of interest (entities), relationships between them and events.

One of the classic IE tasks is NER, which addresses the problem of the identification and classification of predefined concepts (entities). To explain NER we continue by introducing the basic ideas of NLP and ML, and text normalization methods that are used for mapping the same concepts represented in different ways to a standard concept that exists in a domain-specific terminological resource.

2.2.1 Natural language processing

Natural language processing (NLP) is a research area of computer science, artificial intelligence, and computational linguistics, concerned with the interactions between computers and human natural languages. More information about NLP can be found in [65], [66]. NLP works with data represented as unstructured text, which depends on how people express themselves. Text is processed by sentence segmentation and further the segments are

analyzed by applying *tokenization* that is a process of breaking the segments into words that are called tokens. Each of the tokens consists of a string of characters without white space. The tokens can be analyzed by applying *lemmatization* [67] or *stemming* [68], [69]. From linguistics, the lemmatization is the process of grouping together different inflected forms of a word so they can be analyzed as a single item. The uninflected form is called lemma. In computational linguistics, the lemmatization is a process of determining the lemma of a given token (word). It usually works by using vocabulary and morphological analysis of the token in order to return the lemma or dictionary form of the token. Stemming is another approach similar to lemmatization. It usually works by removing the suffixes of the token in order to give a good approximation to the lemma. Further, the tokens can be analyzed by applying *part-of-speech (POS) tagging* that is a process of assigning morphological tags or categories (classes) to each token (e.g., NN (noun, singular or mass), VB (verb, base form), and VBD (verb, past tense)) [70]–[74]. Sometimes it can happen that we are not interested in tokens, but we want to determinate text phrases that are concepts (entities). Chunking [66] is one approach that uses POS tags and identifies short phrases such as noun phrase (NP), verb phrase (VP), preposition phrase (PP), etc. *Chunking* is usually combined with B-I-O tagging scheme, which gives a tag to each token at the beginning of the phrase (B), inside the phrase (I), and outside of any phrase that is tagged (O). For example, the noun phrase (B-NP, I-NP) consists of two tokens, in which the first token is the beginning of the noun phrase and the second token is inside the noun phrase. Despite the morphological analysis, the sentences can be analyzed according to their syntactic structure. The process of working with the syntactic analysis of the sentences is called *parsing* [75].

2.2.2 Machine learning

Machine learning (ML) is a subfield of computer science related to studies of pattern recognition and computational learning theory in artificial intelligence [76]. It focuses on developing algorithms that can learn and make predictions based on data. The data is presented as a training set, which is a collection of instances described by attributes called features. If the training set consists of output labels (classes), given by an expert from the data domain, that are the desired output of the algorithm, then we have *supervised learning*. If the desired output labels are not present in the training set, we have *unsupervised learning*, the goal of which is to find some hidden patterns in the data. Also, *semi-supervised learning* exists, which is a combination of supervised and unsupervised learning. The idea of the ML algorithms is to give further analyses of new unseen instances that are not present in the training set. For example, in supervised learning the output label (class) of these new unseen instances needs to be predicted using the algorithm. Because the ML supervised algorithms perform well for the instances from the training set, their evaluation needs to be done using a test set that consists of instances that are not found in the training set. For this purpose, the training set is often randomly split into two portions, the training set and the test set. Another approach for evaluating the performance of ML supervised algorithms is to use *cross-validation* [77].

ML supervised algorithms are the most used algorithms for information extraction from text. They are based on annotated corpus, which includes text in which the labels of the entities of interest are assigned by domain experts. Using it, different ML models, such as decision trees [78], support vector machines (SVMs) [79], hidden Markov models [80], conditional random fields (CRFs) [81], maximum entropy [82], etc., can be applied in order to achieve better performance. Moreover, the idea of ensemble learning [83] can be used to combine multiple learning algorithms to obtain a better performance that could be obtained from any of the constituent learning algorithms alone.

2.2.3 Named-entity recognition

Named-entity recognition (NER) [17] is a part of information extraction that aims to determine and identify words or phrases in text into predefined labels (classes) that describe concepts of interest in a given domain. Various NER methods exist.

Terminology-driven NER methods, also called dictionary-based NER methods [84]–[86], work by matching the text phrases with concept synonyms that exist in the terminological resources (dictionaries). In order to improve the performance of these methods, instead of strict matching they are combined with some heuristics, such as the generating of words that occur in entity mentions, generating permutations of words in concept synonyms, solving disambiguation problem, etc. The main disadvantage of these methods is that only the entity mentions that exist in the resources will be recognized, but the benefit of using them is related to the frequent updates of the terminological resources with new concepts and synonyms.

Other NER methods are *rule-based NER methods* [87]–[89], which use regular expressions that combine information from terminological resources and characteristics of the entities of interest. The main disadvantage of these methods is the manual construction of the rules, which is a time-consuming task and depends on the domain.

Corpus-based NER methods [22], [23], [90], [91] are based on the evidence that exists in an annotated corpus provided by human experts from the domain and use of ML algorithms to predict the entities labels. These methods are less affected by terminological resources and manual created rules, but the limitation is the existence of an annotated corpus for the domain of interest. The construction of the annotated corpus for a new domain is a time consuming task and requires effort by the human experts to produce it.

To minimize the annotation cost and to exploit unlabelled data in a research related to NER, Active Learning (AL) can be used [92]. It is a semi-supervised learning, or iterative supervised learning, in which an algorithm is able to interactively query the user to obtain the desired outputs at new data points. The examples needed to learn a new concept are chosen by the algorithm and their number can often be much lower than the number of examples required for supervised learning. An active learner uses the unlabelled corpus as the input and generates NER models, while iteratively interacting with the users who annotates sentences queried from the corpus. It usually consists of three components: the annotation interface, the corpus-based NER, and Al component for querying samples. Recently, several studies have also presented the effectiveness of AL to NER tasks [93]–[97].

All previous described NERs rely on pre-defined features which try to capture the specific surface properties of entities, properties of the typical local context, background knowledge, and linguistic information. State-of-the-art NERs rely on the use of the costly handcrafted features and on the output of other NLP tasks. Over the past few years, some recent work on NER have applied Deep Neural Networks (DNNs), which minimize the need of these costly features and have successively advanced the state-of-the-art [98]–[105]. However, this typically requires large amounts of annotated data. To reduce the amount of annotated data, deep learning can be combined with AL [106].

2.2.4 Text normalization

To allow automatic integration of extracted information from a NER task, the information needs to be further processed. The problem that appears is that the same entity can be mentioned in different ways in the same or different documents, using different phrases regarding the text variability. To collect the information for a given entity or even more to combine the information for the entity from different documents, it is crucial to map the entity to a concept that exists in a terminological resource. By mapping it to a concept

from a terminological resource, the extracted entity receives a unique identifier which is the identifier for that entity in the terminological resource. Having unique identifiers helps the process of collecting and combing the information for some entity, even if it has different textual representations. The process of automatic mapping between an entity in text and a concept in a terminological resource is known as text normalization.

A lot of normalization methods are based on *string similarity measures*. String similarity measures give us a metric for similarity (or dissimilarity) between two text strings [107], [108]. They can be performed on a character or term level.

Character-based similarity measures are applied on a character level. Some of them are Longest Common SubString, Damerau-Levenshtein distance [109], Jaro [110], Jaro-Winkler [111], N-gram [112], etc. The Longest Common SubString is based on the length of contiguous chain of characters that exist in both strings. The Damerau-Levenshtein defines the distance between two strings by counting the minimum number of operations needed to transform one string into the other, where an operation is defined as insertion, deletion, or a transposition of two adjacent characters. The Jaro is based on the number and order of the common characters between two strings. The Jaro-Winkler is an extension of Jaro distance that uses a prefix scale which gives more favorable ratings to strings that match from the beginning for a set prefix length. The N-gram is a sub-sequence on n items from a given sentence of text. It can be used to compare the n-grams from each character or word in two strings. Distance is computed by dividing the number of similar n-grams by maximal number of n-grams. Skip-grams are also used, especially in the field of speech processing, where n-grams are formed but in addition to allowing adjacent sequences of words, tokens are skipped [113].

Term-level similarity measures are applied on term (word) level. Some of them are: Manhattan distance [114], Cosine Similarity, Dice's coefficient [115], Euclidean distance, Jaccard similarity [116], Matching coefficient, Overlap coefficient, etc. The Manhattan distance computes the distance that would be travelled to get from one data point to other if a grid-like path is followed. The cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The Dice's coefficient is defined as twice the number of common terms in the compared strings divided by the total number of terms in both strings. The Euclidean distance is the square root of the sum of squared distances between corresponding elements of the two vectors. The Jaccard similarity is computed as the number of shared terms over the number of all unique terms in both strings. The matching coefficient is a vector approach which counts the number of similar terms, on which both vectors are non zero. The overlap coefficient considers two strings a full match if one is a subset of the other.

Hybrid methods also exist, they combine the term-based measures with character-based measures. All above presented similarity measures are based on lexical similarity. For example, words are similar lexically if they have a similar character sequence. Further, these measures can be combined with semantic similarity measures. Words are similar semantically if they have the same thing, are opposite of each other, used in the same way, used in the same context and one is a type of another.

Chapter 3

Deep Statistical Comparison

Nowadays, making a statistical comparison is the essential for comparing the results of a study made using state-of-the-art approaches. Many researchers have problems making a statistical comparison because statistical tools are relatively complex and there are many to chose from. The problem is in selecting the right statistic to apply as a specific measure. For example, researchers often report either the average or median without being aware that averaging is sensitive to outliers and both, the average and median, are sensitive to statistical insignificant differences in the data. Even reporting the standard deviation of the average needs to be made with care since large variances result from the presence of outliers. Furthermore, these statistics only describe the data and do not provide any additional information about the relations that exist between the data. For this, a statistical test needs to be applied. Additionally, the selection of a statistic can influence the outcome of a statistical test. This means that applying the appropriate statistical test requires knowledge of the necessary conditions about the data that must be met in order to apply it. This step is often omitted and researchers simply apply a statistical test, in most cases borrowed from a similar published study, which is inappropriate for their data set. This kind of misunderstanding is all too common in the research community and can be observed in many high-ranking journal papers. In this chapter, we start by explaining the problem definition. We also present an overview of the state-of-the-art approaches used for statistical comparison in classical inferential statistics. We present a novel approach for statistical comparison that provide more robust statistical results. We explicitly point on the difference that exist between practical and statistical significance. Finally, we present the experimental results together with a discussion and the benefits of using the novel proposed approach.

3.1 Problem Definition

Working with experimental data requires statistical analysis, which in most cases involves statistical comparison of data sets that contain the experimental data. Making a statistical comparison, however, requires sufficient knowledge from the user, which includes knowing which conditions must be fulfilled so that the relevant and proper statistical test (e.g., parametric or nonparametric) can be applied [14].

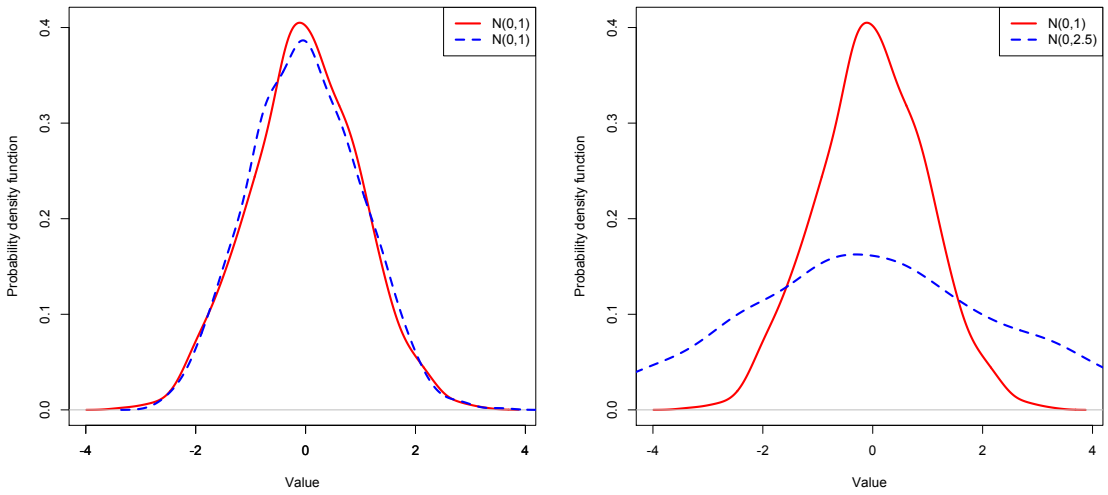
In general, two scenarios are possible. In the first, an experiment is repeated several times on independent samples, so multiple independent measurements are obtained. This is known as a *single-problem analysis* because the samples are from the same type. For example, to measure "beta-carotene" in "apple", multiple "apples" are required (e.g., 15, 20, 25, or more) and the "beta-carotene" parameter is measured using each one. Further, the obtained data set can be compared with other data sets that contain information

for the same parameter from different countries to see if there is a significant statistical difference between the parameter value in those countries. The second scenario is the multiple-problem scenario. An example is when we want to compare data sets from different countries according to a parameter, (e.g., "beta-carotene") with regards to a food group, so instead of using only the multiple measurements obtained on a single food (e.g., "apple"), we also need to include the multiple measurements obtained on other foods that belong to the same food group (e.g., "apricot", "carrot"). In such case, different foods are different problems involved in the comparison and to get a relevant data set, either average or median of the multiple measurements obtained from a single food for each country is typically calculated. This value is a representative value involved in the multiple-problem scenario for a country on that food. More details about these two scenarios that are applicable in any domain are presented in [12], [14]. We call this approach the "common approach" because it is the one most often used for making a statistical comparison.

Statisticians [117] have shown that researchers have difficulties performing empirical studies and this could lead them into misinterpreting their results. Leven et al. [118] state that the hypothesis testing is also frequently misunderstood and abused. To select an appropriate statistical test and to choose between a parametric and a nonparametric test [14], the first step is to check the assumptions of the parametric tests, in other words, the required conditions for the safe use of parametric tests. Further, because the required conditions (normality and homoscedasticity) for the safe use of the parametric statistical tests are usually not satisfied, an appropriate nonparametric statistical test is required. Nonparametric statistical tests are based on a ranking scheme that is used to transform the data prior to analysis. The standard ranking scheme used by many of the nonparametric statistical tests ranks the data sets for each problem (food) separately with the best value being ranked number 1, the second best ranked number 2, and so on. It is assumed that a lower value is better, but this can be changed according to the desire of the user. In case of ties, average ranks are assigned. Unfortunately, using either the average or the median can negatively affect the outcome of the results of a statistical test [25]. For example, averaging is sensitive to outliers, which can be present in the experimental data. For instance, let us suppose that two data sets are given, A_1 and A_2, and the results for 10 measurements are $0, 0, 0, 0, 0, 0, 0, 0, 0, 10$, and $0, 1, 0, 1, 0, 1, 0, 1, 0, 1$, for each data set, respectively. We see that the average of A_1 is 1, while the average of A_2 is 0.5. The standard ranking scheme will rank, A_1 and A_2 as 2 and 1, respectively. From this, it follows that A_2 is better, but we can see that the data set A_1 has only one outlier, which affects the average. In the case when outliers are not present, the average can be in some ϵ-neighborhood, which is defined as the set of all numbers whose distance from a number is less than some specified number ϵ, and the data sets will obtain different ranks. In order to overcome this problem, medians are sometimes used because they are more robust to outliers. However, medians can be in some ϵ-neighborhood, and based on these the data sets will obtain different ranks. The question is, therefore, how to define the ϵ-neighborhood for different problems that have different ranges of obtained data (e.g. 10^{-9}, 10^{-2}, 10^{1}, etc.). Let us suppose that two data sets (e.g., countries), A_1 and A_2, have results from 100 measurements of a parameter on a single food that is distributed according to $N(0; 1)$. Figure 3.1a shows the probability density function of the two data sets. In this case, the distributions are the same, the median values are in some ϵ-neighborhood, and because of this the data sets should obtain the same rank. Now let us suppose that two new data sets, A_1 and A_2, have results from 100 measurements of a parameter on a single food that are distributed according to $N(0; 1)$ and $N(0; 2.5)$, respectively. Figure 3.1b shows the probability density function of the two data sets. In this case, the distributions are not the same and the median values are in some ϵ-neighborhood, and because of this the data sets should obtain different ranks. If

this is the case, then the data sets ranks are obtained either by the averages or the medians, so the data set which has a smaller value for average or median is the better one.

For these reasons we propose a novel approach, which removes the sensitivity of simple statistics to the data and enables the calculation of more robust statistics unaffected by outliers or by errors inside the ϵ-neighborhood. To make things even simpler, we have also created a statistical tool, which automatizes this process, so that even a user with limited statistical knowledge is able to properly apply it to the problem. In our approach, we focus on improving the robustness of current approaches that are used for the statistical comparison.



(a) Same distributions (b) Different distributions

Fig 3.1: Probability density functions.

3.2 Related Work

Statistical analysis is a crucial task, which influences the discussion and the conclusions that are the results of the study. For example, in each scientific published paper in nutrition science, biology, chemistry, physics, computer science, etc., there is a part where the results from a statistical analysis are presented followed by the interpretation of the results. Such statistical analysis involves calculating descriptive statistics to describe the experimental data and statistical comparisons between different data sets in order to explore some relations that exist between the parameters that are measured.

The needs of statistical comparison contributed to a lot of papers and tutorials on how to perform properly the analysis in different domains. In clinical research, there is a study in which the basic principles of hypothesis testing with examples form clinical research studies is presented [119]. The idea of hypothesis testing used with biological and chemical data are presented in [120], [121]. In ML community, Demšar theoretically and empirically examined several suitable statistical tests that can be used for comparing machine learning algorithms [12]. The focus is set on safe and robust nonparametric tests for statistical comparisons of classifiers: the *Wilcoxon signed-rank test* for pairwise comparison and the *Friedman test* with the corresponding post-hoc tests for comparison of more classifiers over

multiple data sets. The results of the tests are also presented with the newly introduced CD (critical difference) diagrams. In the same paper, he gave an experimental setup that can be used to measure the power of the selected statistical tests. Following the statistical tutorial presented for ML algorithms, García et al. [14] presented a study on the use of nonparametric tests for analyzing the evolutionary algorithms' behaviour over optimization problems. The study is conducted in two ways: single-problem analysis and multiple-problem analysis. The single-problem analysis is the scenario when the data comes from multiple runs of the stochastic optimization algorithms on one problem (e.g., test function). This scenario is common in stochastic optimization algorithms, since they are of stochastic nature, meaning we do not have any guarantee that the result will be optimal for every run. Moreover, typically even the path leading to the final solution is often different. So to test the quality of the algorithm, it is not enough to perform just one run, but many of them, from which we can draw some conclusions. The second scenario or the multiple-problem analysis is the scenario when several stochastic optimization algorithms are compared on multiple problems (e.g., test functions). In this case, in most papers the authors use the averaged results for each problem to compose a sample of results for each algorithm. Also, García et al. [15] presented an experimental analysis of power of advanced nonparametric statistical tests for multiple comparisons in design of experiments in computational intelligence and data mining. What is common to all these contributions is that they all use the same statistical methods, two scenarios (single problem and multiple problem) exist, and the only difference is the nature of the data (e.g., experimental food data, biological data, chemical data, performance metric of ML algorithms, global optimum for stochastic optimization algorithm). The idea of this approach regardless of the nature of the data that is used is presented in [122] and we called this approach the "common approach" because it is the most used approach for statistical comparison used in different domains.

3.3 Methodology

Following the recommendations of some papers [12], [14], [122], the results of the common approach can be affected in the case of outliers or some errors inside the ϵ-neighborhood [25], which open a new space for finding more robust statistics that will be used for comparing data sets. To solve problems that arise using the common approach for making a statistical comparison [25], we propose a novel approach called Deep Statistical Comparison (DSC) [29]. We used *deep statistics* because the approach uses a ranking scheme that is based on the whole distribution instead of using averages or medians. It consists of two steps.

The first step is completely new and uses a novel ranking scheme to obtain data that will be further used for statistical comparison. This step is purely nonparametric because it is based on the nonparametric two-sample *Kolmogorov-Smirnov test* (Appendix A). By using it, the ranks of the data sets involved in the comparison are calculated for each problem separately, which are then used for statistical comparison.

The second step is a standard omnibus statistical test that can be either parametric or nonparametric, depending on the required conditions.

3.3.1 The DSC ranking scheme

To define the DSC ranking scheme, let m and k be the number of data sets (e.g., same nutrient parameter from different countries, algorithms, etc.) and the number of problems (e.g., foods, functions, etc.) that are used for making a statistical comparison, respectively,

and n the number of measurements (e.g., independent measurements of the parameter, number of independent runs of an algorithm, etc.) obtained for each data set on the same problem.

Let X_i be an $n \times m$ matrix, where $i = 1, \ldots, k$. The rows of this matrix correspond to the results obtained by multiple measurements on the i-th problem, and the columns correspond to the different data sets involved in the comparison. The matrix element $X_i\,[j, l]$, where $j = 1, \ldots, n$, and $l = 1 \ldots, m$, corresponds to the result obtained by the j-th measurement for the l-th data set.

The ranking scheme is based on the whole distribution, instead of ranking according to the averages or medians. The first step is to compare the probability distributions of multiple measurements for each data set on each problem. For this purpose, the *Kruskal-Wallis test* [123] and the two-sample *Kolmogorov-Smirnov test* can be used.

The *Kruskal-Wallis test* is a nonparametric test for testing whether data samples come from the same distribution [123]. It is a nonparametric alternative to the *One-way ANOVA* and it is used for comparing two or more independent samples or equal or different sample size. The problem of using the *Kruskal-Wallis test* is the misinterpretation of the null hypothesis. The null hypothesis of this test is that the average ranks of the groups are the same. Sometimes, we can see that the null hypothesis is given as "*The samples come from populations with the same distribution.*", which is correct when the samples come from the same distribution and the test will show no difference. However, this result can be misleading because only some kinds of difference in distribution will be detected by the test. For example, if two populations have symmetrical distributions with the same average value, but one is much wider than the other, their distributions are different but the *Kruskal-Wallis test* will not detect any difference between them. Also, the null hypothesis of the *Kruskal-Wallis test* is often said to be that the medians of the groups are the same, but this is only true if one assumes that the shape of the distributions in each group is the same. If the distributions are different, the null hypothesis will be rejected even when the medians are the same.

The two-sample *Kolmogorov-Smirnov test* (KS test) is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare two samples. The two-sample KS test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples (e.g., a difference with respect to location/central tendency, dispersion/variability, skewness, and kurtosis) [124].

Because the *Kruskal-Wallis test* can catch only some kind of difference between the distributions and it is not sensitive to the shape of the distribution, in our case the most appropriate one will be the *two-sample Kolmogorov-Smirnov test*, which is based on the distance between the empirical distribution functions of the two samples. Having in mind the difference that exists between these two tests, in our case the ranking scheme is based on the two-sample KS test.

Let α_{KS} be the significance level used by the two-sample KS test. By using the two-sample KS test, $m \cdot (m-1)/2$ pairwise comparisons between the data sets are performed, and the results are organized in an $m \times m$ matrix, M_i as follows:

$$M_i[p, q] = \begin{cases} p_{\mathrm{value}}, & p \neq q \\ 1, & p = q \end{cases}, \tag{3.1}$$

where p and q are different data sets and $p, q = 1, \ldots, m$.

Because multiple pairwise comparisons are made, this can lead to a *family-wise error rate (FWER)* [125], which is the probability of making one or more false discoveries, or type

I errors, among all hypotheses when performing multiple hypotheses tests. In the cases when fewer data sets are compared, the influence of multiple comparison in the FWER may not be large, but as the number of compared data sets increases, the FWER can increase dramatically. To counteract the problem of multiple comparisons, the *Bonferroni correction* [15] is used to correct the obtained p-values. It is based on the idea of testing u different hypotheses. One way of reducing the FWER is to test each individual hypothesis at a statistical significance level of $\frac{1}{u}$ times the desired maximum significance level. In our case, the number of multiple pairwise comparison, or the number of different hypotheses, is given as $C_m^2 = \binom{m}{2} = \frac{m \cdot (m-1)}{2}$.

Matrix M_i is reflexive and symmetric because $M_i = M_i^T$, but the key point for the ranking scheme is to check transitivity, since the ranking is made according to it. For this purpose, the matrix $M_i^{'}$ is introduced using the following equation:

$$M_i^{'}[p,q] = \begin{cases} 1, & M_i[p,q] \geq \alpha_{\mathrm{KS}}/C_m^2 \\ 0, & M_i[p,q] < \alpha_{\mathrm{KS}}/C_m^2 \end{cases}. \tag{3.2}$$

The elements of the matrix, $M_i^{'}$, are defined according to the p-values obtained by the two-sample KS test corrected by the *Bonferroni correction*. For example if the element $M_i^{'}[p,q]$ is 1, the null hypothesis used in the two-sample KS test, which is the hypothesis that the two data samples obtained by the p-th and the q-th data set come from the same distribution, is not rejected. If the element $M_i^{'}[p,q]$ is 0, then the null hypothesis is rejected, and the two data samples are from different distributions.

Before ranking is performed, the matrix $M_i^{'2}$ is calculated to check transitivity. If $M_i^{'}$ has 1 in each position for which $M_i^{'2}$ has a non-zero element, the transitivity is satisfied, otherwise it is not.

If the transitivity is satisfied, the first step is to split the set of data sets into w disjoint sets of data sets Φ_f, $f = 1, \ldots, w$, so that each data set belongs only in one of these sets. Each of these sets contains the indices of the data sets that are used in the comparison for which the transitivity is satisfied. The cardinality of the union of these sets needs to be m, $\sum_{f=1}^{w} |\Phi_f| = m$. The next step is to define an $w \times 2$ matrix, W_i. The elements of this matrix are defined with the following equation

$$W_i[f,x] = \begin{cases} \mathrm{mean}(\Phi_f\{h\}), & x = 1 \\ |\Phi_f|, & x = 2 \end{cases}, \tag{3.3}$$

where h is the number that is ceiled to the nearest integer of a number obtained by the uniform distribution of a random variable $\mathbf{Y} \sim U(1, |\Phi_f|)$. The data set from each set, whose average value will be used, can be chosen randomly because the data samples for all the data sets that belong to the same set come from the same distribution. The rows of the matrix are then reordered according to the first column which is sorted in ascending order (or descending, depending on the user desire). The $Mean_i$ and C are $w \times 1$ vectors that correspond to the first and the second column of the matrix W_i, respectively. Finally, the ranks to the sets, Φ_f, need to be assigned and organized into an $w \times 1$ vector $Rank_s$. For the set with the lowest average value, $Mean_i[1]$, the rank is defined as

$$Rank_s[1] = \sum_{r=1}^{C[1]} r/C[1]. \tag{3.4}$$

For the remaining sets, the rank is defined as

$$Rank_s[f] = \sum_{r=C[f-1]+1}^{C[f-1]+C[f]} r/C[f]. \tag{3.5}$$

After obtaining the ranks of the sets, each data sets obtains its rank according to the set to which it belongs by using the following equation:

$$Rank[i,l] = Rank_s[f],\ l \in \Phi_f. \tag{3.6}$$

If the transitivity is not satisfied, the first step is to define two $1 \times m$ vectors, $Index_i$ and $Mean_i$, whose elements are the indices of the data sets and the average values of the multiple measurements for each data set. Both vectors are sorted in ascending or descending order according to the average values with regard to the desire of the user. Then, the ranks of the sorted data sets are organized into an $1 \times m$ vector, $Rank_s$, whose elements are defined by the following equation.

$$Rank_s[l] = \begin{cases} l, & \exists!\ Mean_i[l] \in Mean_i \\ \sum_{r=l-c+1}^{l} r/c, & otherwise \end{cases}, \tag{3.7}$$

where c is the number of elements from $Mean_i$ that have a value $Mean_i[l]$. Finally, the data sets obtain their ranks according to the ranks assigned to their average values by using the following equation:

$$Rank[i, Index_i[l]] = Rank_s[l]. \tag{3.8}$$

By using the ranking scheme for the data sets on each problem, an $k \times m$ matrix, $Rank$, is defined. The i-th row of this matrix corresponds to the ranks of the data sets obtained by the ranking scheme using the data samples from the i-th problem. Further, this matrix is used as the input data for making a statistical comparison for multiple-problem analysis.

In general, the new proposed ranking scheme can be used for comparing m different data sets. To be sure that the new ranking scheme gives the data for multiple-problems analysis that is correct, the sum of all ranks obtained for the data sets on each problem needs to be the sum of the arithmetic progression from 1 to m, $\frac{m(m+1)}{2}$. This is necessary because the DSC ranking scheme follows the idea of the fractional ranking scheme used by statistical tests.

The pseudocode of DSC ranking is presented in Algorithm 1.

Algorithm 3.1: Pseudocode of the DSC ranking scheme used to rank data sets on the i-th problem.

```
Calculate the matrix M_i by using Equation 3.1
Recalculate the matrix M_i' by using Equation 3.2
Check the transitivity for the matrix M_i'
if transitivity is satisfied then
    Split the data sets into w disjoint sets Φ_f
    Calculate the matrix W_i by using Equation 3.3
    Obtain the vectors Mean_i and C
    Calculate the vector Rank_s by using Equations 3.4 and 3.5
    Calculate the vector Rank[i,] by using Equation 3.6
else
    Calculate the vector Mean_i and Index_i
    Calculate the vector Rank_s by using Equation 3.7
    Calculate the vector Rank[i,] by using Equation 3.8
end if
return Rank[i,]
```

3.3.2 Omnibus statistical tests

After ranking the data sets, the next step is to choose an appropriate statistical test. The guidelines on which test to choose (parametric or nonparametric) are given in Subsection 2.1.2.

If either the *Friedman test* or *Iman-Davenport test* [14] is used, there is no need to rank the data using the ranking scheme of these tests because the ranks obtained will be the same as the ranks obtained by the DSC ranking scheme. Having the data obtained using the ranking scheme of the DSC, only the test statistic and the p-value need to be calculated in the next step. Using the new ranking scheme we transformed only the data that is available for further analysis, and the test statistic of the appropriate statistical test is not changed.

3.3.3 The difference between practical and statistical significance in data analysis

In industry a crucial task is to consider practical significance. It is, therefore, important to understand the difference between practical significance and statistical significance. For example, let us suppose that we have two measurements for a "beta-carotene" parameter obtained for "apple", one per each country. In the first country, the parameter is detected with order of magnitude of 10^{-10} and in the second country with order of magnitude of 10^{-16}. Between them, a statistical significant difference could be found; however this difference is not significant in a practical sense. To relate industry and research, we need approaches that consider practical significance. In this instance, the data must first be preprocessed at a practical level in which we are interested. The data then needs to be entered into a statistical method in order to find if there is a practical significance between the data sets. The search for practical significance can be made as a preprocessing step. For this purpose, two versions of DSC ranking scheme are introduced. The first one is called sequential. In this version, the order of the measurements can affect the practical significance. To avoid this, a second scheme known as Monte Carlo version is proposed, where a Monte Carlo simulation for each pairwise comparison is made using permutations of the measurements of both data sets. The first ranking scheme can be used for preprocessing of paired samples, and the second one can be used for preprocessing of unpaired samples.

Sequential DSC ranking scheme

In the sequential version of the *DSC* ranking scheme, the search for practical significance is made using a preprocessing step. First, the level of practical significance, ϵ_p, is set by the user. Before comparing the distributions of multiple measurements between each pair of data sets, (p, q), $p, q = 1, \ldots, m$, the data used by each pairwise comparison must be preprocessed in order to consider the practical significance. For each data set on each problem, data is available from n measurements. If the p-th and q-th data sets are involved in a pairwise comparison, their data is preprocessed as follows

$$\begin{cases} p_{\mathrm{g}} = q_{\mathrm{g}} = \frac{p_{\mathrm{g}}+q_{\mathrm{g}}}{2}, & |p_{\mathrm{g}} - q_{\mathrm{g}}| \leq \epsilon_p \\ p_{\mathrm{g}} = p_{\mathrm{g}}, q_{\mathrm{g}} = q_{\mathrm{g}} & |p_{\mathrm{g}} - q_{\mathrm{g}}| > \epsilon_p \end{cases}, \tag{3.9}$$

where $g = 1, \ldots, n$.

The preprocessed data for each pairwise comparison is then used in the statistical test for comparing distributions in order to define the matrix using Equation 3.1.

Monte-Carlo DSC ranking scheme

In the preprocessing step of the sequential version of the DSC ranking scheme, the solutions of two data sets, p and q, for the same problem over measurement g are compared. So the first measurement from the p-th data set is compared with the first measurement of the q-th data set, the second measurement from the p-th data set with the second measurement from the q-th data set, and so on. However, if new n independent measurements for the data sets are obtained, the practical significance may not have the same influence.

To avoid the dependence of the practical significance on the order of the independent measurements, a Monte Carlo [126] version of the *DSC* ranking scheme has been proposed.

In this case, for each n independent measurements of each data set, a different order of the data is defined by generating its permutations. The number of such permutations is $n!$. Because multiple pairwise comparisons are used, for each pairwise comparison, (p, q), $(n!)^2$ different combinations exist, which can be used for checking for practical significance. Each combination has one permutation of the set of multiple measurements from the p-th data set and one permutation from the set of multiple measurements from the q data set.

The ranking scheme randomly selects N different combinations involved in the preprocessing step for each pairwise comparison. For each combination, a search for practical significance is made by preprocessing the data using Equation 3.9. Finally, a statistical test is used to compare distributions for each combination of preprocessed data.

Equation 3.1 needs one p-value for each pairwise comparison that is further corrected to control the FWER, which is the probability of making one or more false discoveries, or type I errors [127], among all hypotheses when performing multiple hypotheses tests.

In our case, by using different selected combinations for the same pairwise comparison we obtain N p-values. Therefore, we need to find a way how to select the appropriate p-value from a set of N p-values. For this, a kernel density estimation [76] is used to estimate the probability density function of the N p-values. The mode of the probability density function is used as an appropriate p-value, which will be used in Equation 3.1 and corrected to control the FWER. The mode of a continuous probability distributions is the value at which its probability density function has its maximum value, so the mode is at the peak.

3.4 Experimental Results

To evaluate the new ranking scheme and to compare it with the results obtained by the common approach [14], we used one-dimensional continuous data. It can be applied on continuous experimental data and it does not depend from the source from where this data comes. Because we are lacking raw experimental food composition data to evaluate the novel proposed approach, we used the results obtained for stochastic optimization algorithms from Black-Box Benchmarking 2015 (BBOB 2015) competition. However, to present the approach using terms form nutrition science, we used countries instead of algorithms, and foods instead of functions. We used the results from 15 countries. For each one, the results for 22 different foods that belong to the same food group were selected. For each country, there are 15 measurements on each food. More details for the data are presented in Appendix B. The 15 measurements are sufficient to show how the proposed novel approach works because the DSC ranking scheme uses the two-sample KS test that can be used for small and large samples, and its corresponding power analysis is presented in [124]. Fifteen measurements are enough to see how the proposed approach works, but if we have more measurements, the approach can only benefit from them, which is the case of each statistical analysis. The literature suggests that the number of independent measurements must be larger (at least 25 or 30), but this number is a result of a central

limit theorem, which states that, given certain conditions, the average of a sufficiently large number of independent random variables, each with well-defined expected value and finite variance, will be approximately normally distributed, regardless of the underlying distribution [128]. A sample size of 25 or 30 is a "large sample" in when discussing "normal approximation". Actually, a sample size of 25 or 30 is useful when you need to choose between the Student's t distribution and normal distribution. For example, a rule of a thumb for using the *Friedman test* is that the number of independent measurements is greater than 10 ($n > 10$) and the number of compared data sets is greater than 5 ($m > 5$) because the Friedman's statistic is distributed according to χ^2 distribution with $m - 1$ degrees of freedom [14]. When the number of independent measurements and data sets is small, the approximation to χ^2 becomes poor and the p-values need to be obtained from tables specially prepared for the *Friedman test* [129], [130].

The statistical comparisons were performed using the *R programming language*. The normality condition was checked by introducing the *Kolmogorov-Smirnov test for normality* that is part of the "*stats*" package [131]. The homoscedasticity was checked by applying the *Levene's test* that is part of the "*lawstat*" package [132]. For the pairwise comparison the parametric *paired t-test* and the alternative nonparametric *Wilcoxon signed-rank test* were used that are part of the "*stats*" package, while for comparison of more than two data sets the parametric *repeated-measures ANOVA* from the "*stats*" package, and the nonparametric alternatives, the *Friedman test*, *Friedman aligned-ranks test*, and *Iman-Davenport test* from the "*scmamp*" package [133] were used.

Let Ψ be the set of 15 countries, $C_i, i = 1, \ldots, 15$, selected for testing. We have 22 different problems (foods), f_i, $i = 1, .., n$, that belong to the same food group, F.

We perform six experiments. In the first and the second one we consider a multiple comparisons scenario among three and four data sets, respectively. Three and four data sets were selected because when the number of compared data sets is smaller, the benefits of using DSC are more observable. In the third experiment multiple comparisons with a control data set are analyzed. The fourth and fifth experiments give the results of using the sequential and Monte-Carlo DSC ranking scheme that considers a practical significance, respectively. In the sixth experiment, the scenario of all pairwise comparisons is presented.

3.4.1 Multiple comparisons among three data sets

One hundred random subsets of 3 distinct countries of Ψ were generated (combinations without repetition) and used for statistical comparisons. For each combination, a sample of results for each country is composed following the common approach by averaging the data from multiple measurements for each problem for each country. After obtaining the data for statistical comparison over multiple problems, the next step is to select an appropriate statistical test that will be used to make a statistical comparison between the selected countries. In order to select between a parametric and a nonparametric test, the required conditions for the safe use of the parametric tests are checked, and the appropriate statistical test is selected accordingly. To check the normality, a *Kolmogorov-Smirnov test for normality* is used, while to check the homoscedasticity, the *Levene's test* based on averages is used.

In our case the statistical comparisons are performed between three countries, so if the required conditions for the safe use of the parametric tests are satisfied, a *repeated-measures ANOVA* is used, but if they are not satisfied, statistical comparisons are performed by using one of the following alternative nonparametric tests: *Friedman test*, *Friedman algined-ranks test*, or *Iman-Davenport test*. Using the *Friedman test* in the case when the number of countries is less than 5, the approximation of the χ^2 statistic becomes poor and therefore the p-values are obtained from specially prepared tables [129], [130].

Instead of averaging the data from multiple measurements for each problem of each country, the DSC ranking scheme is used to obtain the ranks of the data sets for each problem separately. Having the ranking data for the countries for each problem, the data is ready to be used for statistical comparison. As in the common approach, first the required conditions for the safe use of parametric tests are checked, and according to the result an appropriate statistical test is selected. In our case, the significance level that is used for the ranking scheme is set to $\alpha_{\mathrm{KS}} = 0.05$, and the significance level that is used in the statistical test is set to $\alpha = 0.05$.

In Table 3.1 the p-values for 25 combinations out of 100 generated are selected and presented in order to show the difference between the common approach using averages and the DSC approach. The p_{valueF} and p_{valueID} correspond to the p-values obtained by the *Friedman test* and the *Iman-Davenport test*, respectively, and they are reported for both approaches. From this table, we can see that the results for the first 15 combinations (1-15) differ. By using the common approach with averages, the null hypothesis is rejected, so there is a significant statistical difference between the parameter value of those countries, while when using DSC, the null hypothesis is not rejected, so we can assume that there is no significant statistical difference between the countries. The number of such combinations in our experiment is 15 out of 100. For the next 5 combinations (16-20), both approaches give the same results, the p-values obtained are greater than 0.05, so the null hypothesis is not rejected, and we can assume that there is no significant statistical difference between the parameter value of those countries. The number of such combinations in our experiment is 11 out of 100, from which 5 randomly selected ones are presented in the table. For the last 5 combinations from this table (21-25), the results obtained for both approaches are the same, the p-values are smaller than the significance level, so the null hypothesis is rejected, and we can assume that there is a significant statistical difference between the countries. The number of such combinations in our experiment is 74 out of 100, from which 5 are randomly selected and presented in the table.

To explain the difference that appears using both approaches, one example of each group of combinations is randomly selected and presented in detail. For the first 15 combinations, the 10th combination was selected, where the statistical comparison between C_2, C_4, and C_{15} was made. First, the common approach with averages is introduced. Using it to obtain data for multiple-problem analysis, the average of the results from multiple measurements for each problem for each country is calculated. Using the 15 measurements for each problem for each country, the data for these three countries that is composed for making a statistical comparisons is presented in Table 3.2.

The results obtained by the *Friedman test* and *Iman-Davenport test* show that the null hypothesis is rejected, and we can assume that there is a significant statistical difference between the parameter values of these countries. In order to obtain a better visual inspection, the distribution of the data for each country obtained by the common approach with averages is presented in Figure 3.2. The data is presented on a base 10 logarithmic scale, and because of this the 1^{st} and 5^{th} rows from Table 3.2 are omitted. By omitting them the probability distributions of the data do not change because they provide the same values for the 1^{st} problem and the 5^{th} problem. Based on visual inspection, one may assume that there is no difference between the distributions of the data of the three countries. To confirm this, a two-sample KS test was used to compare the pairs of countries, and the p-values are presented in Table 3.3. The obtained p-values are greater than 0.05, so the null hypothesis is not rejected in all three cases, therefore the distributions of data between the pairs of countries are the same.

The data obtained by the DSC ranking scheme for the same countries is given in Table 3.4, and is further used for statistical comparison. First, the required conditions for the safe

Table 3.1: Statistical comparisons of three countries.

	Countries	Common approach		DSC approach	
		p_{value_F}	$p_{value_{ID}}$	p_{value_F}	$p_{value_{ID}}$
1	C_{11}, C_{14}, C_1	*(.00)	*(.00)	(.28)	(.28)
2	C_2, C_{13}, C_1	*(.03)	*(.02)	(.79)	(.79)
3	C_{14}, C_9, C_1	*(.02)	*(.02)	(.55)	(.56)
4	C_1, C_{11}, C_{13}	*(.02)	*(.02)	(.31)	(.32)
5	C_4, C_{13}, C_{15}	*(.04)	*(.03)	(.80)	(.81)
6	C_{11}, C_{14}, C_2	*(.00)	*(.00)	(.07)	(.07)
7	C_1, C_3, C_2	*(.03)	*(.03)	(.64)	(.64)
8	C_2, C_{11}, C_{15}	*(.01)	*(.01)	(.18)	(.18)
9	C_9, C_1, C_{15}	*(.04)	*(.04)	(.42)	(.43)
10	C_2, C_4, C_{15}	*(.03)	*(.02)	(.98)	(.98)
11	C_{11}, C_2, C_{13}	*(.02)	*(.02)	(.17)	(.17)
12	C_{14}, C_8, C_9	*(.04)	*(.04)	(.21)	(.21)
13	C_{15}, C_1, C_2	*(.03)	*(.03)	(.86)	(.87)
14	C_{14}, C_9, C_{11}	*(.03)	*(.03)	(.11)	(.11)
15	C_4, C_{13}, C_{14}	*(.02)	*(.02)	(.77)	(.74)
16	C_2, C_8, C_3	(.19)	(.19)	(.70)	(.71)
17	C_8, C_2, C_{13}	(.24)	(.24)	(.86)	(.87)
18	C_3, C_{14}, C_8	(.25)	(.26)	(.97)	(.97)
19	C_3, C_{14}, C_{13}	(.06)	(.06)	(.86)	(.87)
20	C_{11}, C_2, C_{13}	(.42)	(.43)	(.17)	(.17)
21	C_{11}, C_8, C_{15}	*(.00)	*(.00)	*(.02)	*(.02)
22	C_1, C_{15}, C_5	*(.00)	*(.00)	*(.00)	*(.00)
23	C_4, C_1, C_7	*(.00)	*(.00)	*(.01)	*(.01)
24	C_{14}, C_7, C_3	*(.00)	*(.00)	*(.00)	*(.00)
25	C_1, C_3, C_6	*(.00)	*(.00)	*(.00)	*(.00)

* indicates that the null hypothesis is rejected, using $\alpha = 0.05$

p_{valueF} corresponds to the p-value obtained by the *Friedman test*

p_{valueID} corresponds to the p-value obtained by the *Iman-Davenport test*

use of the parametric tests are checked, but the condition for normality is not satisfied, so the appropriate statistical tests are the *Friedman test* and the *Iman-Davenport test*. The p-values obtained by these two tests are greater than 0.05, so the null hypothesis is not rejected, and we can assume that there is no significant statistical difference between the parameter value of these three countries.

The ranks obtained by the *Friedman test* using the common approach with averages for the same countries are presented in Table 3.5. Comparing them with the ranks obtained by the DSC ranking scheme presented in Table 3.4, the difference between the ranks that appears using both approaches can be clearly observed. To explain the difference that appears between the ranks obtained by both approaches, separate problems (foods) are discussed in detail.

Figure 3.3 shows the cumulative distributions (step functions) and the average values (vertical lines) obtained from the multiple measurements of f_3 for the three countries. The ranks obtained by the common approach with averages are 1.00, 2.00, and 3.00, and they

Table 3.2: Averages for multiple problem analysis obtained by the common approach for the countries: C_2, C_4, and C_{15}.

F	$\mathbf{C_2}$	$\mathbf{C_4}$	$\mathbf{C_{15}}$
f_1	0.00000e+00	0.00000e+00	7.10511e-09
f_2	3.37974e-09	1.74196e-09	6.20692e-09
f_3	1.29934e-09	1.70852e-09	7.65295e-09
f_4	4.78763e-09	3.95106e-09	2.84506e-09
f_5	-2.13163e-15	-2.13163e-15	-2.13163e-15
f_6	1.70953e+01	1.99725e+01	4.37864e+00
f_7	1.29522e+01	1.15325e+01	6.90837e+00
f_8	1.42791e+00	1.49952e+00	1.68230e+00
f_9	7.16261e+00	6.59892e+00	5.87362e+00
f_{10}	2.74150e+04	2.44383e+04	1.04656e+04
f_{11}	8.05797e+01	8.77173e+01	8.33617e+01
f_{12}	7.09055e+01	1.79466e+01	4.12477e+03
f_{13}	2.82980e+00	2.88984e+00	1.01204e+00
f_{14}	2.84126e-03	2.50924e-03	2.25311e-03
f_{15}	8.33959e+01	7.63882e+01	5.77356e+01
f_{16}	4.55429e+00	5.10948e+00	2.61043e+00
f_{17}	5.64192e+00	5.51401e+00	4.45787e+00
f_{18}	1.71806e+01	1.76836e+01	1.03145e+01
f_{19}	1.64144e+00	1.54069e+00	1.08443e+00
f_{20}	9.11419e-01	8.29647e-01	8.21958e-01
f_{21}	8.23275e+00	7.79717e+00	5.75912e+00
f_{22}	1.61024e+01	1.65596e+01	1.26976e+01

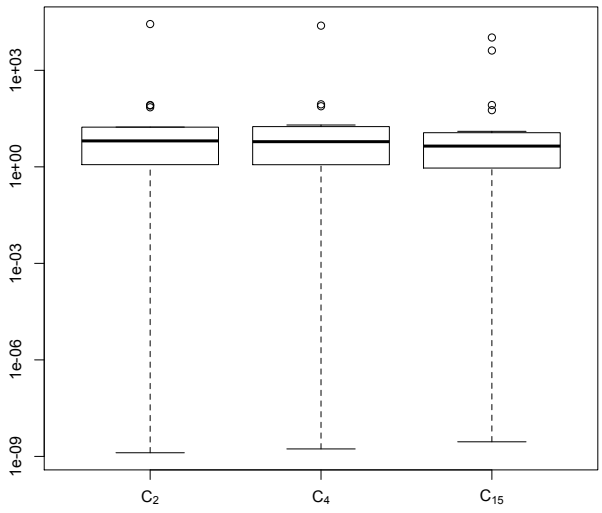


Fig 3.2: Boxplot of the data for the countries C_2, C_4, and C_{15}.

are different because all of them have a different average value. Also, it is noticeable that the

Table 3.3: Two-sample *Kolmogorov-Smirnov test* between the pairs of countries.

p-value	**(C_2, C_4)**	**(C_2, C_{15})**	**(C_4, C_{15})**
p_{value}	1	(.86)	(.99)

Table 3.4: Ranks obtained by the DSC ranking scheme for the countries C_2, C_4, and C_{15}.

F	**C_2**	**C_4**	**C_{15}**
f_1	1.50	1.50	3.00
f_2	1.50	1.50	3.00
f_3	1.50	1.50	3.00
f_4	2.00	2.00	2.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	2.00	2.00	2.00
f_8	2.00	2.00	2.00
f_9	2.00	2.00	2.00
f_{10}	2.00	2.00	2.00
f_{11}	2.00	2.00	2.00
f_{12}	2.00	2.00	2.00
f_{13}	2.00	3.00	1.00
f_{14}	2.00	2.00	2.00
f_{15}	2.00	2.00	2.00
f_{16}	2.00	2.00	2.00
f_{17}	2.00	2.00	2.00
f_{18}	2.00	2.00	2.00
f_{19}	3.00	2.00	1.00
f_{20}	2.00	2.00	2.00
f_{21}	2.00	2.00	2.00
f_{22}	2.00	2.00	2.00

differences between the averages are in some ϵ-neighborhood, $e-09$. The ranks obtained by the DSC ranking scheme for the same problem are 1.50, 1.50, and 3.00. The DSC ranking scheme uses the cumulative distributions to rank the countries. From the figure, one may assume that there is no significant difference between the cumulative distributions of C_2 and C_4, but it is obvious that they differ from the cumulative distribution of C_{15}. This result is also obtained by using a two-sample KS test where the p-values obtained for the pairs of countries are 0.94 (C_2, C_4), 0.00 (C_2, C_{15}), and 0.00 (C_4, C_{15}). Because multiple pairwise comparisons are made, these p-values are corrected using the *Bonfferoni correction*. In this case, the transitivity of the matrix $M_3{}'$ is satisfied, so the set of all countries is split into two distinct sets {C_2, C_4}, and {C_{15}}, and the ranks obtained are defined by Equations 3.4 and 3.5. The difference is that the DSC approach assigns ranks according to the information obtained by the whole distribution. If the distributions are the same, then there is no need to rank the countries according to their averages that can be affected by the presence of outliers or the ϵ-neighborhood.

In Figure 3.4, the cumulative distributions (step functions) and the average values (vertical lines) obtained from multiple measurements of f_{19} for the three countries are presented. The ranks obtained using the common approach with averages are 3.00, 2.00, and 1.00, and are different because all of them have different average values. Also, it is

Table 3.5: Ranks obtained by the *Friedman test* using the common approach with averages for the countries C_2, C_4, and C_{15}.

F	$\mathbf{C_2}$	$\mathbf{C_4}$	$\mathbf{C_{15}}$
f_1	1.50	1.50	3.00
f_2	2.00	1.00	3.00
f_3	1.00	2.00	3.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	3.00	2.00	1.00
f_8	1.00	2.00	3.00
f_9	3.00	2.00	1.00
f_{10}	3.00	2.00	1.00
f_{11}	1.00	3.00	2.00
f_{12}	2.00	1.00	3.00
f_{13}	2.00	3.00	1.00
f_{14}	3.00	2.00	1.00
f_{15}	3.00	2.00	1.00
f_{16}	2.00	3.00	1.00
f_{17}	3.00	2.00	1.00
f_{18}	2.00	3.00	1.00
f_{19}	3.00	2.00	1.00
f_{20}	3.00	2.00	1.00
f_{21}	3.00	2.00	1.00
f_{22}	2.00	3.00	1.00

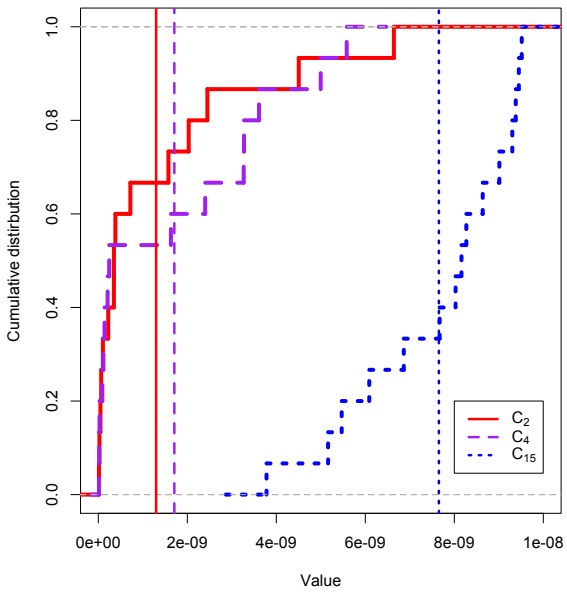


Fig 3.3: Cumulative distributions (step functions) and average values (vertical lines) for f_3 of C_2, C_4, and C_{15}.

noticeable that the differences between the averages are not in some small ϵ-neighborhood. The ranks obtained by the DSC ranking scheme for the same problem are 3.00, 2.00, and 1.00. From the figure, it is not clear if there is a significant difference between the cumulative distributions of C_2, C_4, and C_{15}. To check this, a two-sample KS test was used and the p-values obtained for the pairs of countries are 0.68 (C_2, C_4), 0.01 (C_2, C_{15}), and 0.03 (C_4, C_{15}). Because multiple pairwise comparisons are made, these p-values are further corrected using the *Bonfferoni correction* and compared with the significance level 0.02. In this case, the transitivity of the matrix $M_{19}{}'$ is not satisfied because the data from C_2 and C_4 comes from the same distribution, the data from C_4 and C_{15} comes from the same distribution, but the data from C_2 and C_{15} is not from the same distribution. In this case the ranks obtained are defined by Equation 3.7, and are the same as the ranks obtained using the common approach with averages. This result is reasonable since the averages are not affected by the ϵ-neighborhood.

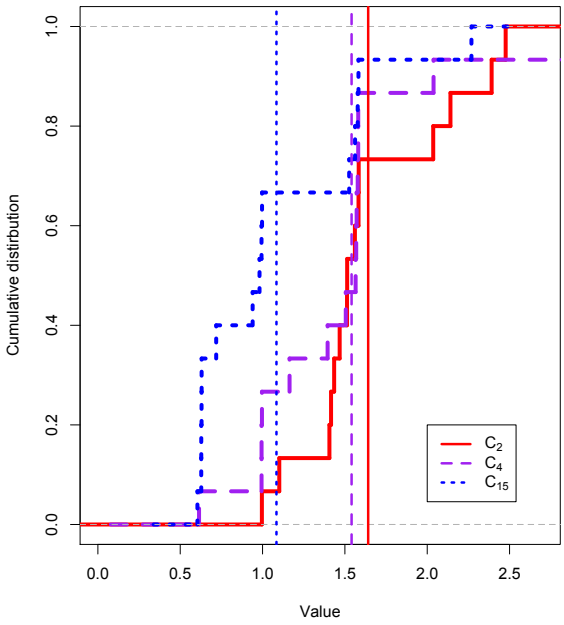


Fig 3.4: Cumulative distributions (step functions) and average values (vertical lines) for f_{19} of C_2, C_4, and C_{15}.

In Figures 3.5 and 3.6 the cumulative distributions (step functions) and the average values (vertical lines) obtained from the multiple measurements of f_{21} and f_{22} for the three countries are presented, respectively. In this case only f_{21} will be explained in detail because the same scenario happens when f_{22} is analyzed. The ranks obtained using the common approach with averages are 3.00, 2.00, and 1.00, and they are different because all of them have different average values. Also, it is noticeable that the differences between the averages are in some reasonable ϵ-neighborhood. The ranks obtained by the DSC ranking scheme for the same problem are 2.00, 2.00, and 2.00. From the figure, one may assume that there is no significant difference between the cumulative distributions of C_2, C_4, and C_{15}. This result is also obtained by using the two-sample KS test and the p-values obtained for the pairs of countries are 0.93 (C_2, C_4), 0.68 (C_2, C_{15}), and 0.94 (C_4, C_{15}). Because multiple pairwise comparisons are made, these p-values are further corrected using the *Bonfferoni correction*. In this case, the transitivity of the matrix $M_{21}{}'$ is satisfied, but the set of countries is not split into distinct sets because all of the countries belong to one

set, {C_2, C_4, C_{15}}, and the ranks obtained are the ranks defined by Equations 3.4 and 3.5. By using the DSC, the different ranks obtained when the averages are affected by outliers or ϵ-neighborhood can be omitted.

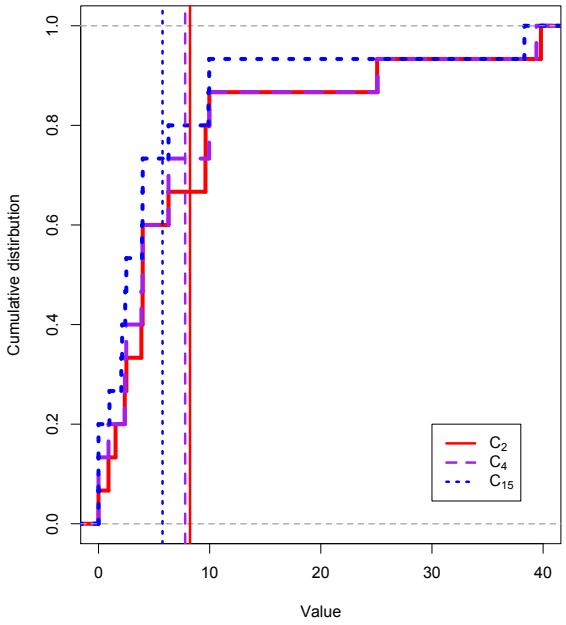


Fig 3.5: Cumulative distributions (step functions) and average values (vertical lines) for f_{21} of C_2, C_4, and C_{15}.

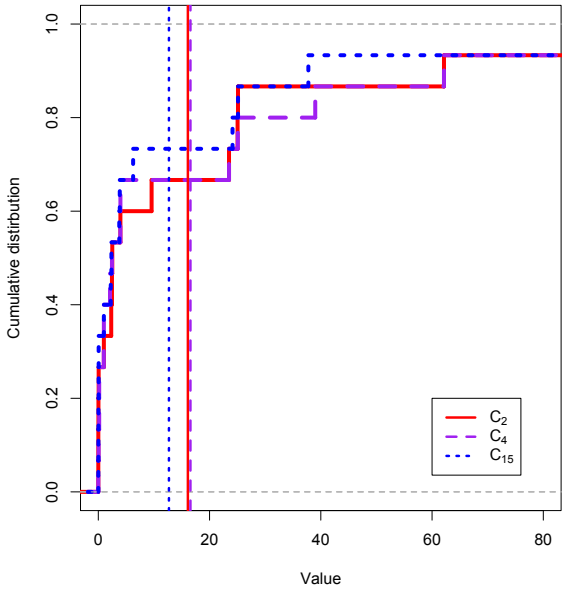


Fig 3.6: Cumulative distributions (step functions) and average values (vertical lines) for f_{22} of C_2, C_4, and C_{15}.

Because there are differences between the results obtained by the common approach

with averages and the DSC approach, the result is also compared with a modified version of the common approach, where instead of average, the median from the multiple measurements for each problem of each country is used (Table 3.6). This is made because the median is less sensitive to outliers than average value. Using the common approach with medians the p-values obtained by the *Friedman test* and the *Iman-Davenport test* are $p_{\text{value}F} = 0.01$ and $p_{\text{value}ID} = 0.01$, respectively. Accordingly, the null hypothesis is rejected, so there is a significant statistical difference between the countries. So even when the median is used, there is still a significant statistical difference. In such case, this difference is caused by the ranking scheme that is influenced by the differences between the medians that are in some ϵ-neighborhood and consequently the ranks are 1.00, 2.00, and 3.00.

Table 3.6: Ranks obtained by the *Friedman test* using the common approach with medians for the countries C_2, C_4, and C_{15}.

F	$\mathbf{C_2}$	$\mathbf{C_4}$	$\mathbf{C_{15}}$
f_1	1.50	1.50	3.00
f_2	2.00	1.00	3.00
f_3	2.00	1.00	3.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	3.00	2.00	1.00
f_8	2.00	1.00	3.00
f_9	3.00	2.00	1.00
f_{10}	2.00	3.00	1.00
f_{11}	2.00	3.00	1.00
f_{12}	3.00	1.00	2.00
f_{13}	2.00	3.00	1.00
f_{14}	3.00	2.00	1.00
f_{15}	3.00	2.00	1.00
f_{16}	2.00	3.00	1.00
f_{17}	3.00	2.00	1.00
f_{18}	2.00	3.00	1.00
f_{19}	2.00	3.00	1.00
f_{20}	3.00	1.00	2.00
f_{21}	3.00	2.00	1.00
f_{22}	3.00	2.00	1.00

Then another combination of three countries from the second group was randomly selected. The countries are C_8, C_2, and C_{13} (17th combination presented in Table 3.1). Using both approaches the results from the *Friedman test*, $p_{\text{value}F} = 0.24$ (common approach with averages) and $p_{\text{value}F} = 0.86$ (DSC approach), and *Iman-Davenport test*, $p_{\text{value}ID} = 0.24$ (common approach with averages) and $p_{\text{value}ID} = 0.87$ (DSC approach), fail to reject the null hypothesis, so there is no significant statistical difference between the countries. The ranks obtained by the *Friedman test* using the common approach for these countries and the ranks obtained by the DSC ranking scheme are presented in Table 3.7. Figure 3.7 shows a box plot of the average data values of each country. The data is presented on base 10 logarithmic scale, and because of this, the values for the 1^{st} and 5^{th} problems are omitted. Figure 3.7 indicates that there is no difference between the distributions of the

data for each country, which is the same result obtained by both approaches, the common approach with averages and the DSC approach.

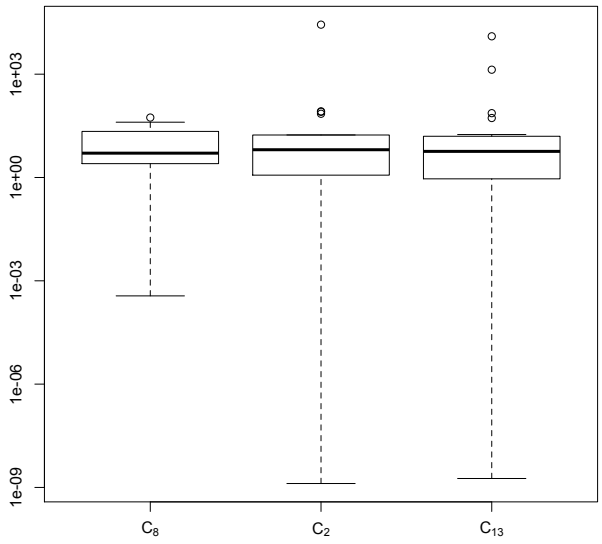


Fig 3.7: Boxplot of the data for the countries C_8, C_2, and C_{13}.

At the end of the first experiment, a combination of countries from the third group was randomly selected. The countries are C_1, C_3, and C_6, which is the 25^{th} combination presented in Table 3.8. Because the required conditions for the safe use of the parametric tests are not satisfied, the nonparametric *Friedman test*, $p_{\text{value}F} = 0.00$ (common approach with averages) and $p_{\text{value}F} = 0.00$ (DSC approach), and *Iman-Davenport test*, $p_{\text{value}ID} = 0.00$ (common approach with averages) and $p_{\text{value}ID} = 0.00$ (DSC approach), were used. The results reveal a significant statistical difference between the countries. To show the difference between both approaches, the ranks obtained by the *Friedman test* using the common approach and the ranks obtained by the DSC ranking scheme are presented in Table 3.8. In Figure 3.8 the box plot of the average data values for each country is presented. The data is presented on base 10 logarithmic scale, and due to the logarithmic properties, the values for the 1^{st} and 5^{th} problems are omitted. By using the graphical representation of the probability distributions, one may assume that there is a difference between the distributions of the data of the countries, which is the result obtained using both, the common approach with averages and the DSC approach.

Also, we need to note here that when comparing more than two countries over multiple problems if the required conditions for the safe use of the parametric tests are not satisfied, the alternative nonparametric test is *Friedman aligned-ranks test*. This test can be used when the number of countries being compared is small and comparability among the problems is desirable. The p-values obtained by the *Friedman aligned-rank test* for three combinations of countries, one per each group, from the first experiment, are presented in Table 3.9.

The benefit of using the DSC ranking scheme is that it is based on the whole distribution, instead of using only one statistic, i.e. as the average or median in the common approach. By using the DSC ranking scheme to obtain data for multiple-problem analysis, outliers that can affect averaging are avoided because whole distribution of data from

Table 3.7: Ranks for the countries C_8, C_2, and C_{13}.

(a) *Friedman test*

F	$\mathbf{C_8}$	$\mathbf{C_2}$	$\mathbf{C_{13}}$
f_1	3.00	1.00	2.00
f_2	3.00	1.00	2.00
f_3	3.00	1.00	2.00
f_4	3.00	2.00	1.00
f_5	3.00	1.50	1.50
f_6	1.00	3.00	2.00
f_7	1.00	3.00	2.00
f_8	3.00	2.00	1.00
f_9	1.00	3.00	2.00
f_{10}	1.00	3.00	2.00
f_{11}	1.00	3.00	2.00
f_{12}	1.00	2.00	3.00
f_{13}	3.00	2.00	1.00
f_{14}	1.00	2.00	3.00
f_{15}	1.00	3.00	2.00
f_{16}	2.00	3.00	1.00
f_{17}	1.00	3.00	2.00
f_{18}	1.00	3.00	2.00
f_{19}	3.00	2.00	1.00
f_{20}	3.00	2.00	1.00
f_{21}	1.00	3.00	2.00
f_{22}	1.00	2.00	3.00

(b) DSC ranking scheme

F	$\mathbf{C_8}$	$\mathbf{C_2}$	$\mathbf{C_{13}}$
f_1	2.50	1.00	2.50
f_2	3.00	1.50	1.50
f_3	3.00	1.00	2.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	1.00	3.00	2.00
f_7	1.00	2.50	2.50
f_8	3.00	1.50	1.50
f_9	2.00	2.00	2.00
f_{10}	1.00	2.50	2.50
f_{11}	1.00	2.50	2.50
f_{12}	1.00	2.50	2.50
f_{13}	2.00	2.00	2.00
f_{14}	1.00	2.00	3.00
f_{15}	1.50	3.00	1.50
f_{16}	2.00	2.00	2.00
f_{17}	1.00	2.50	2.50
f_{18}	1.00	2.50	2.50
f_{19}	3.00	2.00	1.00
f_{20}	3.00	1.50	1.50
f_{21}	2.00	2.00	2.00
f_{22}	2.00	2.00	2.00

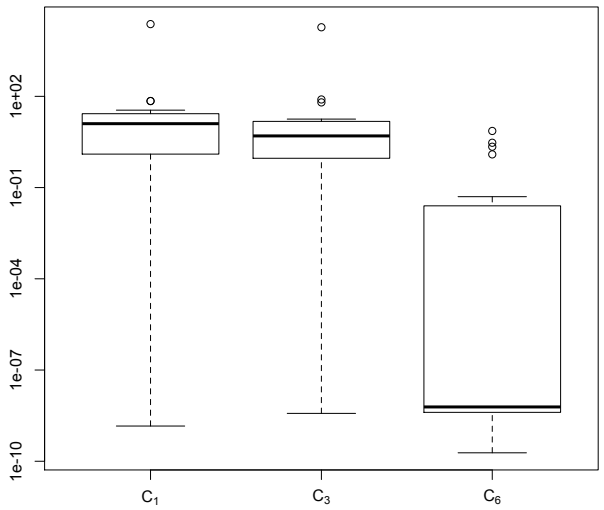


Fig 3.8: Boxplot of the data for the countries C_1, C_3, and C_6.

Table 3.8: Ranks for the countries C_1, C_3, and C_6.

(a) *Friedman test*

F	$\mathbf{C}_1$	$\mathbf{C}_3$	$\mathbf{C}_6$
f_1	1.50	1.50	3.00
f_2	1.00	2.00	3.00
f_3	2.00	3.00	1.00
f_4	1.00	2.00	3.00
f_5	2.00	2.00	2.00
f_6	3.00	2.00	1.00
f_7	2.00	3.00	1.00
f_8	3.00	2.00	1.00
f_9	3.00	2.00	1.00
f_{10}	3.00	2.00	1.00
f_{11}	3.00	2.00	1.00
f_{12}	3.00	2.00	1.00
f_{13}	3.00	2.00	1.00
f_{14}	3.00	2.00	1.00
f_{15}	2.00	3.00	1.00
f_{16}	3.00	2.00	1.00
f_{17}	3.00	2.00	1.00
f_{18}	3.00	2.00	1.00
f_{19}	3.00	2.00	1.00
f_{20}	2.00	1.00	3.00
f_{21}	3.00	2.00	1.00
f_{22}	3.00	2.00	1.00

(b) DSC ranking scheme

F	$\mathbf{C}_1$	$\mathbf{C}_3$	$\mathbf{C}_6$
f_1	1.50	1.50	3.00
f_2	1.00	2.00	3.00
f_3	2.00	3.00	1.00
f_4	1.50	1.50	3.00
f_5	2.00	2.00	2.00
f_6	2.50	2.50	1.00
f_7	2.50	2.50	1.00
f_8	3.00	2.00	1.00
f_9	2.50	2.50	1.00
f_{10}	2.50	2.50	1.00
f_{11}	2.50	2.50	1.00
f_{12}	3.00	2.00	1.00
f_{13}	3.00	2.00	1.00
f_{14}	3.00	2.00	1.00
f_{15}	2.50	2.50	1.00
f_{16}	2.50	2.50	1.00
f_{17}	2.50	2.50	1.00
f_{18}	2.50	2.50	1.00
f_{19}	2.50	2.50	1.00
f_{20}	1.50	1.50	3.00
f_{21}	2.50	2.50	1.00
f_{22}	2.00	2.00	2.00

multiple measurements for each problem of each country is used. In addition the DSC approach gives more robust results even when the common approach is used with the median values instead of average values. However in both cases, using the common approach with either averages or medians, the compared countries will obtain different ranks even if the differences between the medians are in some ϵ-neighborhood, and this can lead to wrong conclusions.

Table 3.9: Statistical comparisons of three countries using the *Friedman aligned-ranks test.*

	Countries	**Common approach**	**DSC approach**
		p_{valueFa}	p_{valueFa}
1	C_2, C_4, C_{15}	$^*(.00)$	(.96)
2	C_2, C_{13}, C_1	(.37)	(.63)
3	C_1, C_3, C_6	$^*(.00)$	$^*(.00)$

* indicates that the null hypothesis is rejected, using $\alpha = 0.05$

p_{valueFa} corresponds to the p-value obtained by the *Friedman aligned-ranks test*

3.4.2 Multiple comparisons among four data sets

Again one hundred subsets for four distinct countries were selected and used for statistical comparisons. The results of 6 randomly selected combinations of countries are presented in Table 3.10. For the first two combinations (1-2) both approaches provided different results. Using the common approach with averages the null hypothesis is rejected, while using DSC the null hypothesis is not rejected. The number of such combinations is 7 out of 100. For the next two combinations (3-4) both approaches provided the same result, i.e. the null hypothesis is not rejected. The number of such combinations in our experiment is 7 out of 100. For the last two combinations (5-6) the results provided by both approaches are the same and the null hypothesis is rejected. The number of such combinations in our experiment is 86 out of 100. In Figures 3.9, 3.10, and 3.11, box plots of the average data values for each country are presented for the 1^{st}, 3^{rd}, and 5^{th} combination from Table 3.10. The data is presented on base 10 logarithmic scale, and because of this, the values obtained for the 1^{st} and 5^{th} problems are omitted.

As in our previous experiment, by omitting them the probability distributions of the data are not changed because they provided the same values for these two problems. Based on a visual inspection of the figures, the results obtained by the DSC approach are valid.

Table 3.10: Statistical comparisons of four countries.

	Countries	**Common approach**		**DSC approach**	
		$p_{value\,F}$	$p_{value\,ID}$	$p_{value\,F}$	$p_{value\,ID}$
1	C_3, C_1, C_{15}, C_8	$^{*}(.04)$	$^{*}(.03)$	(.59)	(.60)
2	C_3, C_{13}, C_1, C_{15}	$^{*}(.01)$	$^{*}(.01)$	(.55)	(.56)
3	C_1, C_{11}, C_2, C_9	(.60)	(.60)	(.46)	(.47)
4	C_{15}, C_3, C_8, C_9	(.06)	(.06)	(.34)	(.34)
5	C_1, C_6, C_4, C_3	$^{*}(.00)$	$^{*}(.00)$	$^{*}(.00)$	$^{*}(.00)$
6	C_6, C_{11}, C_9, C_3	$^{*}(.00)$	$^{*}(.00)$	$^{*}(.00)$	$^{*}(.00)$

* indicates that the null hypothesis is rejected, using $\alpha = 0.05$
$p_{value\,F}$ corresponds to the p-value obtained by the *Friedman test*
$p_{value\,ID}$ corresponds to the p-value obtained by the *Iman-Davenport test*

3.4.3 Multiple comparisons with a control data set

In the previous two experiments, we made comparisons using three and four countries. First, the data for statistical analysis was obtained by the DSC ranking scheme, which is based on comparing distributions, and then the obtained data is used as input data for some appropriate statistical test. The results from the previous experiments showed that the omnibus statistical tests in combination with the DSC approach provide more robust results than the common approach when the results are affected by outliers or by a misleading ranking scheme. However, it can happen that more than three countries are compared against one another, as multiple comparisons with a control country. In this case, the data for the statistical analysis is obtained using the DSC ranking scheme and then used as the input for an appropriate omnibus statistical test. Usually some nonparametric statistical test is selected, such as the *Friedman test*, the *Friedman aligned-ranks test*, or the *Iman-Davenport test*, since they allow a comparison of more than 2 countries. If the null hypothesis is not rejected, then there is no significant statistical difference between the countries and we do not need to perform any post-hoc tests. If

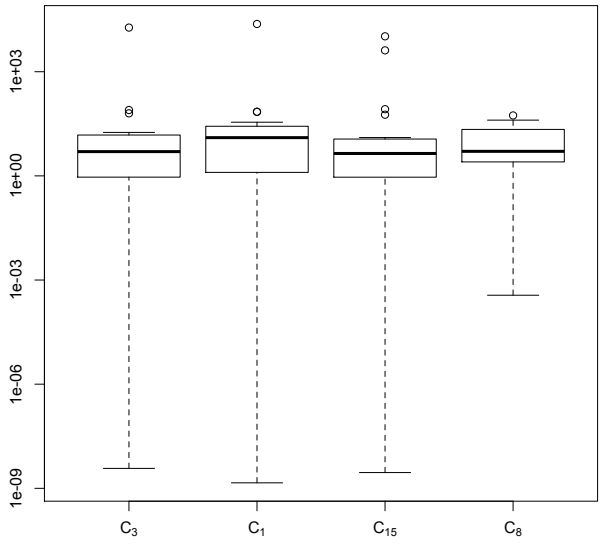


Fig 3.9: Boxplot of the data for the countries C_3, C_1, C_{15}, and C_8.

the null hypothesis is rejected, then we should continue with post-hoc tests, for which we can use different procedures such as *Bonferroni-Dunn procedure* [59], *Holm procedure* [60], *Hochberg procedure* [61], etc. To see what happens when the DSC approach involves more than four countries we presented an experiment in which 10 countries are selected and one of them is compared against the other 9 countries. First, the input data is obtained using the DSC ranking scheme, which is then used by an appropriate omnibus statistical test. In our experiment the compared countries are: C_7, C_{10}, C_1, C_8, C_{12}, C_{14}, C_4, C_3, C_5, and C_{15}. The ranks obtained by the *Friedman test* for the common approach with averages are presented in Table 3.11, while the ranks obtained using the DSC ranking scheme are presented in Table 3.12. From these two tables we can see that the ranks obtained by both approaches differ. The p-values obtained by both approaches with the *Friedman test* are 0.00, so the null hypothesis is rejected and there is a significant statistical difference in the countries.

To continue with multiple comparisons with a control country, a suitable post-hoc test needs to be selected. In this case, a family of hypotheses can be defined, all related to the control country, and we can obtain a p-value that determines the degree of rejection for each hypothesis. The test statistic for comparing the ith and jth country depends on the omnibus statistical test. In our case, for the *Friedman test*, the test statistic is defined as

$$z = (R_i - R_j)/\sqrt{\frac{m(m+1)}{6k}}, \tag{3.10}$$

where R_i and R_j are the average ranks obtained by the *Friedman test* of the compared countries, m is the number of compared countries, and k is the number of different problems (foods). The value of z is used to find the probability from the table of the standardized normal distribution, $N(0, 1)$, which is then compared with the significance level α. The post-hoc tests differ in the way they adjust the value of α to compensate for multiple comparisons. Different procedures can be used for p-value adjustment such as *Bonferroni-Dunn procedure*, *Holm procedure*, *Hochberg procedure*, etc. For example, to compare C_5

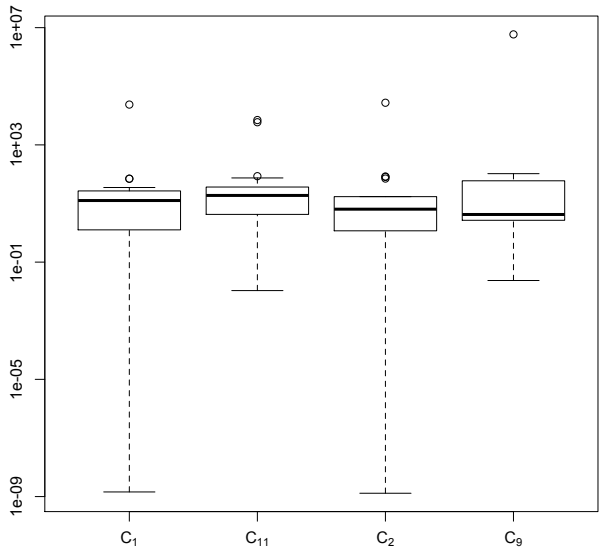


Fig 3.10: Boxplot of the data for the countries C_1, C_{11}, C_2, and C_9.

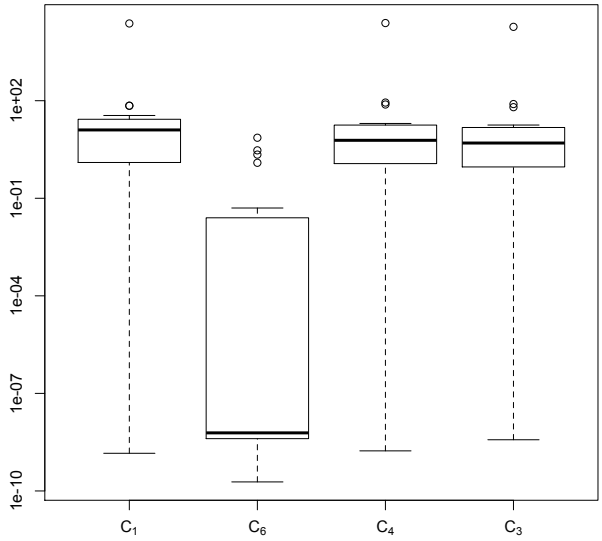


Fig 3.11: Boxplot of the data for the countries C_1, C_6, C_4, and C_3.

Table 3.11: Ranks obtained by the *Friedman test* using the common approach with averages for 10 countries.

F	$\mathbf{C_7}$	$\mathbf{C_{10}}$	$\mathbf{C_1}$	$\mathbf{C_8}$	$\mathbf{C_{12}}$	$\mathbf{C_{14}}$	$\mathbf{C_4}$	$\mathbf{C_3}$	$\mathbf{C_5}$	$\mathbf{C_{15}}$
f_1	4.00	9.00	2.00	8.00	10.00	6.00	2.00	2.00	5.00	7.00
f_2	4.00	9.00	2.00	8.00	10.00	5.00	1.00	3.00	6.00	7.00
f_3	6.00	9.00	1.00	8.00	10.00	5.00	2.00	3.00	7.00	4.00
f_4	7.00	9.00	4.00	8.00	10.00	1.00	3.00	5.00	6.00	2.00
f_5	4.50	4.50	4.50	9.00	10.00	4.50	4.50	4.50	4.50	4.50
f_6	2.00	9.00	8.00	5.00	10.00	3.00	7.00	6.00	1.00	4.00
f_7	1.00	10.00	7.00	3.00	9.00	5.00	6.00	8.00	2.00	4.00
f_8	2.00	9.00	7.00	8.00	10.00	4.00	5.00	3.00	1.00	6.00
f_9	2.00	9.00	8.00	4.00	10.00	7.00	5.00	6.00	1.00	3.00
f_{10}	2.00	9.00	7.00	3.00	10.00	5.00	8.00	6.00	1.00	4.00
f_{11}	2.00	10.00	6.00	3.00	9.00	5.00	8.00	4.00	1.00	2.00
f_{12}	2.00	9.00	6.00	4.00	10.00	7.00	5.00	3.00	1.00	8.00
f_{13}	2.00	8.00	9.00	7.00	10.00	4.00	6.00	5.00	1.00	3.00
f_{14}	2.00	9.00	8.00	3.00	10.00	6.00	5.00	7.00	1.00	4.00
f_{15}	1.00	6.00	8.00	3.00	7.00	4.00	9.00	10.00	2.00	5.00
f_{16}	2.00	10.00	7.00	5.00	9.00	4.00	8.00	6.00	1.00	3.00
f_{17}	1.00	5.00	9.00	3.00	4.00	6.00	10.00	8.00	2.00	7.00
f_{18}	1.00	7.00	9.00	3.00	5.00	6.00	10.00	8.00	2.00	4.00
f_{19}	2.00	10.00	7.00	8.00	9.00	4.00	6.00	5.00	1.00	3.00
f_{20}	3.00	9.00	7.00	8.00	10.00	2.00	5.00	6.00	1.00	4.00
f_{21}	2.00	3.00	10.00	4.00	9.00	6.00	8.00	7.00	1.00	5.00
f_{22}	2.00	5.00	10.00	3.00	6.00	4.00	9.00	8.00	1.00	7.00

with the other 9 countries, the z statistic for each pair between C_5 and one of the other 9 countries (9 pairs) is calculated using Equation 3.10. The p-value is obtained using the standardized normal distribution. Then, a post-hoc procedure for adjusting the p-values is selected to control the FWER. In our case, we selected the *Holm* and *Hochberg* procedures because they are more powerful than the *Bonferroni-Dunn procedure*. The p-values obtained by both post-hoc procedures for the common approach with averages are presented in Table 3.13, while the p-values obtained by both post-hoc procedures for the DSC approach are presented in Table 3.14. From these tables, we can see that both post-hoc procedures give the same results by using $\alpha = 0.05$, the first eight null hypotheses are rejected, so there is a significant statistical difference between C_5 and each of the first eight countries presented in those tables. Also, for both approaches, both post-hoc procedures fail to reject the last null hypothesis, so there is no significant statistical difference between C_5 and C_7.

To see the influence of an increasing number of countries on the DSC ranking scheme, let us look at the problem f_{22} from Table 3.12. The DSC ranking scheme is based on comparing distributions, so multiple pairwise comparisons are made and then the obtained p-values are corrected using the *Bonferroni correction* in order to control the FWER. However, the correction comes at the cost of increasing the probability of producing false negatives, or reducing the statistical power [15]. This is a general criticism, which can be applied to any FWER control approach and it is not specific only to the *Bonferroni correction*. To see the difference that exists if we do not correct the p-values by the *Bonferroni correction*, the ranks obtained by the DSC ranking scheme without controlling

Table 3.12: Ranks obtained by the DSC ranking scheme for 10 countries.

F	$\mathbf{C_7}$	$\mathbf{C_{10}}$	$\mathbf{C_1}$	$\mathbf{C_8}$	$\mathbf{C_{12}}$	$\mathbf{C_{14}}$	$\mathbf{C_4}$	$\mathbf{C_3}$	$\mathbf{C_5}$	$\mathbf{C_{15}}$
f_1	6.00	9.00	2.00	6.00	10.00	6.00	2.00	2.00	6.00	6.00
f_2	4.00	9.00	2.00	8.00	10.00	5.00	1.00	3.00	6.00	7.00
f_3	6.00	9.00	1.00	8.00	10.00	5.00	2.00	3.00	7.00	4.00
f_4	7.00	9.00	4.00	8.00	10.00	1.00	3.00	5.00	6.00	2.00
f_5	5.50	5.50	5.50	5.50	5.50	5.50	5.50	5.50	5.50	5.50
f_6	2.00	9.00	8.00	5.00	10.00	3.00	7.00	6.00	1.00	4.00
f_7	1.50	7.00	7.00	3.00	7.00	7.00	7.00	7.00	1.50	7.00
f_8	2.00	9.00	7.00	8.00	10.00	4.00	5.00	3.00	1.00	6.00
f_9	2.00	9.00	8.00	4.00	10.00	7.00	5.00	6.00	1.00	3.00
f_{10}	2.00	9.00	7.00	3.00	10.00	5.00	8.00	6.00	1.00	4.00
f_{11}	2.00	10.00	6.00	3.00	9.00	5.00	8.00	4.00	1.00	2.00
f_{12}	2.00	9.00	6.00	4.00	10.00	7.00	5.00	3.00	1.00	8.00
f_{13}	2.00	8.00	9.00	7.00	10.00	4.00	6.00	5.00	1.00	3.00
f_{14}	1.50	9.50	8.00	3.00	9.50	5.50	5.50	5.50	1.50	5.50
f_{15}	1.00	6.00	8.00	3.00	7.00	4.00	9.00	10.00	2.00	5.00
f_{16}	2.00	10.00	7.00	5.00	9.00	4.00	8.00	6.00	1.00	3.00
f_{17}	1.00	5.00	9.00	3.00	4.00	6.00	10.00	8.00	2.00	7.00
f_{18}	1.50	7.00	7.00	3.00	7.00	7.00	7.00	7.00	1.50	7.00
f_{19}	2.00	10.00	7.00	8.00	9.00	4.00	6.00	5.00	1.00	3.00
f_{20}	4.50	9.00	4.50	9.00	9.00	4.50	4.50	4.50	1.00	4.50
f_{21}	2.00	3.00	10.00	4.00	9.00	6.00	8.00	7.00	1.00	5.00
f_{22}	5.50	5.50	5.50	5.50	5.50	5.50	5.50	5.50	5.50	5.50

Table 3.13: Multiple comparisons with a control country (C_5) by using the average ranks obtained by the *Friedman test* with averages.

j	$\mathbf{C_5}$ **vs.**	**z**	**Unadjusted** p_{value}	**Holm** p_{value}	**Hochberg** p_{value}
1	C_{12}	-7.3444	1.0329e-13	9.2965e-13	9.2965e-13
2	C_{10}	-6.3734	9.2381e-11	7.3905e-10	7.3905e-10
3	C_1	-4.8299	6.8294e-07	4.7806e-06	4.7806e-06
4	C_4	-4.1328	1.7917e-05	1.0750e-04	1.0750e-04
5	C_3	-3.6846	1.1449e-04	5.7247e-04	5.7247e-04
6	C_8	-3.41084	3.2384e-04	1.2953e-03	1.2953e-03
7	C_{15}	-2.7884	2.6484e-03	7.9452e-03	7.1705e-03
8	C_{14}	-2.6888	3.5852e-03	7.9452e-03	7.1705e-03
9	C_7	-0.3485	3.637133e-01	3.6371e-01	3.6371e-01

the FWER are presented in Table 3.15. From it, we can see that the ranks of the countries for the problem f_{22} differ when the *Bonferroni correction* is not applied. So, when the number of countries that are involved in the comparison increases, the correction can influence the ranks that will be obtained. In this case, the significance level that is used for the *Kolmogorov-Smirnov test* is $\alpha_{\text{KS}} = 0.05$. Then in order to control the FWER, each of the multiple pairwise hypotheses is tested with a significance level α_{KS}/u, where u is the number of hypotheses, which in our case is $\binom{10}{2} = 45$. So, each hypothesis that is involved in the multiple pairwise comparisons is tested with a significance level $0.05/45 = 0.0011$. Because of this, the ranks that are obtained by using the *Bonferroni correction* are equal.

Table 3.14: Multiple comparisons with a control country (C_5) by using the average ranks obtained by the DSC ranking scheme.

j	**C_5 vs.**	**z**	**Unadjusted p_{value}**	**Holm p_{value}**	**Hochberg p_{value}**
1	C_{12}	-6.7222	8.9592e-12	8.0633e-11	8.0633e-11
2	C_{10}	-6.025	8.4582e-10	6.7665e-09	6.7665e-09
3	C_1	-4.1328	1.7917e-05	1.2542e-04	1.2542e-04
4	C_4	-3.6099	1.5310e-04	9.1862e-04	9.8162e-04
5	C_3	-3.0622	1.0983e-03	5.4916e-03	5.1825e-03
6	C_8	-3.0124	1.2956e-03	5.4916e-03	5.1825e-03
7	C_{15}	-2.8133	2.4517e-03	7.3553e-03	5.7183e-03
8	C_{14}	-2.7635	2.8591e-03	7.3553e-03	5.7183e-03
9	C_7	-0.4730	3.1809e-01	3.1809e-01	3.1809e-01

When we compare fewer countries, it is less likely that the correction will influence the result, while when the number of countries increases, the correction influence on the result is emphasized. To avoid the influence of the correction used in the DSC ranking scheme, an appropriate way of comparing multiple comparisons with a control country by using the DSC ranking scheme without the influence of the correction is to perform multiple *Wilcoxon* tests for each pairwise comparison. In this case, input data is obtained by the DSC ranking scheme, and the *Bonferroni correction* does not influence the result because we have only one hypothesis involved, $u = 1$. Since the comparison involves comparing distributions between two countries, the hypothesis is tested at a significance level $0.05/1 = 0.05$. The p-values obtained by the *Wilcoxon test* for each hypothesis are presented in Table 3.16. The *Wilcoxon test* performs a comparison between two data sets. The p-values in a pairwise comparison are independent from each other. If we try to extract a conclusion by using more pairwise comparisons made by the *Wilcoxon test*, we will accumulate error coming from the combination of pairwise comparisons, and consequently we are losing the control of the FWER. The true statistical significance [14] for combining pairwise comparisons is given by

$$\begin{aligned} p_{value} &= P(Reject\ H_0 | H_0\ true) \\ &= 1 - P(Fail\ to\ reject\ H_0 | H_0\ true) \\ &= 1 - P(Fail\ to\ reject\ A_k = A_i, i = 1, \ldots, k-1 | H_0\ true) \\ &= 1 - \prod_{i=1}^{k-1} P(Fail\ to\ reject\ A_k = A_i | H_0\ true) \\ &= 1 - \prod_{i=1}^{k-1} [1 - P(Reject\ A_k = A_i | H_0\ true)] \\ &= 1 - \prod_{i=1}^{k-1} [1 - p_{value_{H_i}}]. \end{aligned} \tag{3.11}$$

From Table 3.16, before the FWER is not controlled, the conclusion that C_5 is a significant statistical different than C_{10}, C_1, C_8, C_{12}, C_{14}, C_4, C_3 and C_{15}, with a significance level $\alpha = 0.05$ can not be correct. C_5 is a significant statistical different than each of the 8 countries considering independent pairwise comparisons because the obtained p-values are smaller than $\alpha = 0.05$. The true statistical difference for combining pairwise comparisons for these 8 hypotheses can be obtained using Equation 3.11, which in our case is 0.0173,

Table 3.15: Ranks obtained by the DSC ranking scheme without FWER control for 10 countries.

F	$\mathbf{C_7}$	$\mathbf{C_{10}}$	$\mathbf{C_1}$	$\mathbf{C_8}$	$\mathbf{C_{12}}$	$\mathbf{C_{14}}$	$\mathbf{C_4}$	$\mathbf{C_3}$	$\mathbf{C_5}$	$\mathbf{C_{15}}$
f_1	4.00	9.00	2.00	8.00	10.00	6.00	2.00	2.00	5.00	7.00
f_2	4.00	9.00	2.00	8.00	10.00	5.00	1.00	3.00	6.00	7.00
f_3	6.00	9.00	1.00	8.00	10.00	5.00	2.00	3.00	7.00	4.00
f_4	7.00	9.00	4.00	8.00	10.00	1.00	3.00	5.00	6.00	2.00
f_5	5.50	5.50	5.50	5.50	5.50	5.50	5.50	5.50	5.50	5.50
f_6	2.00	9.00	8.00	5.00	10.00	3.00	7.00	6.00	1.00	4.00
f_7	1.50	7.00	7.00	3.00	7.00	7.00	7.00	7.00	1.50	7.00
f_8	2.00	9.00	7.00	8.00	10.00	4.00	5.00	3.00	1.00	6.00
f_9	2.00	9.00	8.00	4.00	10.00	7.00	5.00	6.00	1.00	3.00
f_{10}	2.00	9.00	7.00	3.00	10.00	5.00	8.00	6.00	1.00	4.00
f_{11}	2.00	10.00	6.00	3.00	9.00	5.00	8.00	4.00	1.00	2.00
f_{12}	2.00	9.00	6.00	4.00	10.00	7.00	5.00	3.00	1.00	8.00
f_{13}	2.00	8.00	9.00	7.00	10.00	4.00	6.00	5.00	1.00	3.00
f_{14}	1.50	9.00	8.00	3.00	10.00	5.50	5.50	5.50	1.50	5.50
f_{15}	1.00	6.00	8.00	3.00	7.00	4.00	9.00	10.00	2.00	5.00
f_{16}	2.00	10.00	7.00	5.00	9.00	4.00	8.00	6.00	1.00	3.00
f_{17}	1.00	5.00	9.00	3.00	4.00	6.00	10.00	8.00	2.00	7.00
f_{18}	1.00	7.00	9.00	3.00	5.00	6.00	10.00	8.00	2.00	4.00
f_{19}	2.00	10.00	7.00	8.00	9.00	4.00	6.00	5.00	1.00	3.00
f_{20}	3.00	9.00	7.00	8.00	10.00	2.00	5.00	6.00	1.00	4.00
f_{21}	2.00	3.00	10.00	4.00	9.00	6.00	8.00	7.00	1.00	5.00
f_{22}	2.00	5.00	10.00	3.00	6.00	4.00	9.00	8.00	1.00	7.00

Table 3.16: Multiple comparisons with a control country (C_5) by using multiple *Wilcoxon* tests with the DSC ranking scheme.

j	$\mathbf{C_5}$ **vs.**	p_{value}
1	C_7	9.9999e-01
2	C_{10}	5.0490e-06
3	C_1	4.8480e-03
4	C_8	1.4510e-05
5	C_{12}	5.0490e-06
6	C_{14}	6.3010e-04
7	C_4	7.7680e-03
8	C_3	3.0820e-03
9	C_{15}	1.0570e-03

and it is smaller than the significance level used, $\alpha = 0.05$. From this, we can conclude that C_5 is a significant statistical different than these 8 countries.

At the end if we compare the results of the post-hoc tests using the DSC ranking scheme with the *Friedman test* and the multiple pairwise *Wilcoxon* tests, we find that for this experiment the results are the same. This happens because in this example, in the case of the *Friedman test*, only four problems (f_{14}, f_{18}, f_{20}, and f_{22}) are affected by the *Bonferroni correction*. In general if more problems are affected by the correction, the results from the *Friedman test* and the multiple pairwise *Wilcoxon* tests can be different.

From this experiment, we can conclude that the number of compared data sets is a crucial factor that can affect the performance of the DSC ranking scheme. When the number of compared data sets increases, the DSC ranking scheme is affected by the correction used to control the FWER. To solve this problem, when the number of compared data sets increases it is better to use multiple *Wilcoxon tests*, one per each pairwise comparison, and then to combine the obtained p-values in order to calculate the true p-value of the comparison.

3.4.4 Multiple comparisons among three data sets by using the sequential DSC ranking scheme

In the forth experiment, the focus is on practical significance. Here, a DSC approach is made using the sequential version of the DSC ranking scheme. The practical significance level, ϵ_p, is set by the user. Table 3.17, shows the DSC results corresponding to a significance level 0.05 for the *DSC* ranking scheme and a significance level $\alpha = 0.05$ for the selected omnibus statistical test. The results are presented for different practical significance levels.

Table 3.17: Statistical comparison of three countries using DSC sequential ranking scheme.

Countries		ϵ_p							
		10^{-9}	10^{-6}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
1	C_{11}, C_{14}, C_1	1	1	1	1	1	1	1	1
2	C_2, C_{13}, C_1	1	1	1	1	1	1	1	1
3	C_{14}, C_9, C_1	1	1	1	1	1	1	1	1
4	C_1, C_{11}, C_{13}	1	1	1	1	1	1	1	1
5	C_4, C_{13}, C_{15}	1	1	1	1	1	1	1	1
6	C_{11}, C_{14}, C_2	1	1	1	1	1	1	1	1
7	C_1, C_3, C_2	1	1	1	1	1	1	1	1
8	C_2, C_{11}, C_{15}	1	1	1	1	1	1	1	1
9	C_9, C_1, C_{15}	1	1	1	1	1	1	1	1
10	C_2, C_4, C_{15}	1	1	1	1	1	1	1	1
11	C_{11}, C_2, C_{13}	1	1	1	1	1	1	1	1
12	C_{14}, C_8, C_9	1	1	1	1	1	1	1	1
13	C_{15}, C_1, C_2	1	1	1	1	1	1	1	1
14	C_{14}, C_9, C_{11}	1	1	1	1	1	1	1	1
15	C_4, C_{13}, C_{14}	1	1	1	1	1	1	1	1
16	C_2, C_8, C_3	1	1	1	1	1	1	1	1
17	C_8, C_2, C_{13}	1	1	1	1	1	1	1	1
18	C_3, C_{15}, C_8	1	1	1	1	1	1	1	1
19	C_3, C_{14}, C_{13}	1	1	1	1	1	1	1	1
20	C_{11}, C_2, C_{13}	1	1	1	1	1	1	1	1
21	C_{11}, C_8, C_{15}	0	0	0	1	1	0	1	1
22	C_1, C_{15}, C_5	0	0	0	0	0	0	1	1
23	C_4, C_1, C_7	0	0	0	0	0	0	1	1
24	C_{14}, C_7, C_3	0	0	0	0	0	0	1	1
25	C_1, C_3, C_6	0	0	0	0	0	0	1	1

0 - indicates that the null hypothesis fails to reject, $p_{value} < 0.05$
1 - indicates that the null hypothesis fails to reject, $p_{value} \geq 0.05$
p_{value} corresponds to the p-value obtained by the *Friedman test*

From Tables 3.17, it is obvious that the practical significance of the compared countries is different for different practical significance levels. When the practical significance level

increases, combinations exist where the practical significance level differs, but above which there is no difference between the countries. Also, there are combinations in which the practical significance changes non periodically between 0 and 1, or between difference and no difference. To see what happens in the case of these combinations, let us look at the 21^{st} combination in Table 3.17, and practical significance levels, $\epsilon_p \in \{10^{-1}, 10^0, 10^1\}$. This combination is shown in detail in Table 3.18, where ranks from the sequential version of the DSC ranking scheme are shown for the selected values of the practical significance level. From this table, we see that the ranks for different practical significance levels differ.

Table 3.18: Ranks for the countries C_{11}, C_8, and C_{15}, using the first version of the *DSC* ranking scheme.

(a) $\epsilon_p = 10^{-1}$

F	$\mathbf{C}_{11}$	$\mathbf{C}_8$	$\mathbf{C}_{15}$
f_1	2.00	2.00	2.00
f_2	3.00	2.00	1.00
f_3	2.50	2.50	1.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	3.00	1.50	1.50
f_7	2.50	1.00	2.50
f_8	3.00	2.00	1.00
f_9	3.00	1.50	1.50
f_{10}	2.50	1.00	2.50
f_{11}	2.50	1.00	2.50
f_{12}	2.00	2.00	2.00
f_{13}	3.00	2.00	1.00
f_{14}	2.00	2.00	2.00
f_{15}	2.00	2.00	2.00
f_{16}	3.00	2.00	1.00
f_{17}	2.00	1.00	3.00
f_{18}	1.50	1.50	3.00
f_{19}	2.00	3.00	1.00
f_{20}	2.50	2.50	1.00
f_{21}	2.00	2.00	2.00
f_{22}	2.00	2.00	2.00

(b) $\epsilon_p = 10^0$

F	$\mathbf{C}_{11}$	$\mathbf{C}_8$	$\mathbf{C}_{15}$
f_1	2.00	2.00	2.00
f_2	3.00	2.00	1.00
f_3	2.50	2.50	1.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	3.00	1.50	1.50
f_7	2.50	1.00	2.50
f_8	3.00	2.00	1.00
f_9	3.00	1.50	1.50
f_{10}	2.50	1.00	2.50
f_{11}	2.50	1.00	2.50
f_{12}	2.00	2.00	2.00
f_{13}	3.00	2.00	1.00
f_{14}	2.00	2.00	2.00
f_{15}	2.00	2.00	2.00
f_{16}	3.00	2.00	1.00
f_{17}	2.50	1.00	2.50
f_{18}	1.50	1.50	3.00
f_{19}	2.00	3.00	1.00
f_{20}	2.50	2.50	1.00
f_{21}	2.00	2.00	2.00
f_{22}	2.00	2.00	2.00

(c) $\epsilon_p = 10^1$

F	$\mathbf{C}_{11}$	$\mathbf{C}_8$	$\mathbf{C}_{15}$
f_1	2.00	2.00	2.00
f_2	3.00	2.00	1.00
f_3	2.50	2.50	1.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	3.00	1.50	1.50
f_7	2.00	2.00	2.00
f_8	2.00	2.00	2.00
f_9	2.00	2.00	2.00
f_{10}	2.50	1.00	2.50
f_{11}	2.50	1.00	2.50
f_{12}	2.00	2.00	2.00
f_{13}	2.00	2.00	2.00
f_{14}	2.00	2.00	2.00
f_{15}	2.00	2.00	2.00
f_{16}	3.00	2.00	1.00
f_{17}	2.00	2.00	2.00
f_{18}	2.00	2.00	2.00
f_{19}	2.00	3.00	1.00
f_{20}	2.00	2.00	2.00
f_{21}	2.00	2.00	2.00
f_{22}	2.00	2.00	2.00

For example, the ranks obtained for $\epsilon_p = 10^{-1}$ and $\epsilon_p = 10^0$ differ only in one problem, f_{17}, but this is enough to affect the *Friedman* test statistic, which is an appropriate omnibus statistical test for these comparison. The test statistic in both cases is 5.614 and 6.295, with p-values 0.060 and 0.043, respectively. When the significance level is set to 0.05, $\alpha = 0.05$, the result reveal that there is no significant practical difference between the parameter of the compared countries when $\epsilon_p = 10^{-1}$, while the result is the opposite when $\epsilon_p = 10^0$. In both cases, the p-values are around the significance level that we are interested in.

When the practical significance level increases to $\epsilon_p = 10^1$, the ranks differ from those when $\epsilon_p = 10^{-1}$ and $\epsilon_p = 10^0$ in 8 problems. The test statistic is 1.386 with a p-value 0.5, so there is no significant practical difference between the countries.

To see what happens with this combination at a single problem level, we focus on problem f_{17}. For this problem, the ranks differ for different values of practical significance. This is because comparisons of distributions are made to obtain the ranks and as a result of preprocessing the distributions of the data change. To explain this, the results for the problem f_{17} are presented in detail. In Figure 3.12, the cumulative distributions for pairwise comparisons between the three countries C_{11}, C_8, and C_{15}, are given for different practical significance levels. The distributions change for different practical significance

levels.

When the practical significance level is $\epsilon_p = 10^{-1}$, the ranks are 2.00, 1.00, and 3.00, respectively. The p-values from the two-sample KS test are 0.000 for (C_{11},C_8), 0.007 for (C_{11},C_{15}), and 0.000 for (C_8,C_{15}). Each hypothesis is tested at a statistical significance of 0.016, taking into account the *Bonferroni* correction. In this example, all distributions differ (figures 6.1a, 6.1b, and 6.1c), the transitivity is not satisfied and the countries obtain their ranks according to their average values.

When the practical significance level increases to $\epsilon_p = 10^0$, the rankings are 2.50, 1.00, 2.50, respectively. In this case, the p-values for the pairwise comparisons are 0.002 for (C_{11},C_8), 0.028 for (C_{11},C_{15}), and 0.000 for (C_8,C_{15}), which are further tested at a significance level, 0.016, taking into account the *Bonferroni* correction. From the p-values, the distributions between the countries C_{11} and C_{15} are the same and the transitivity is satisfied, so that the set of all countries is split on two sets {C_{11}, C_{15}} and {C_8}, and the countries obtain a rank from the set to which they belong.

When the practical significance level increases to $\epsilon_p = 10^1$, the rankings are 2.00, 2.00, and 2.00, respectively. The p-values are 1.000 for (C_{11},C_8), 1.000 for (C_{11},C_{15}), and 1.000 for (C_8,C_{15}). From the obtained p-values and from the figures 6.1g, 6.1h, and 6.1i, we see that the distributions are the same, so the countries are all received the same rank.

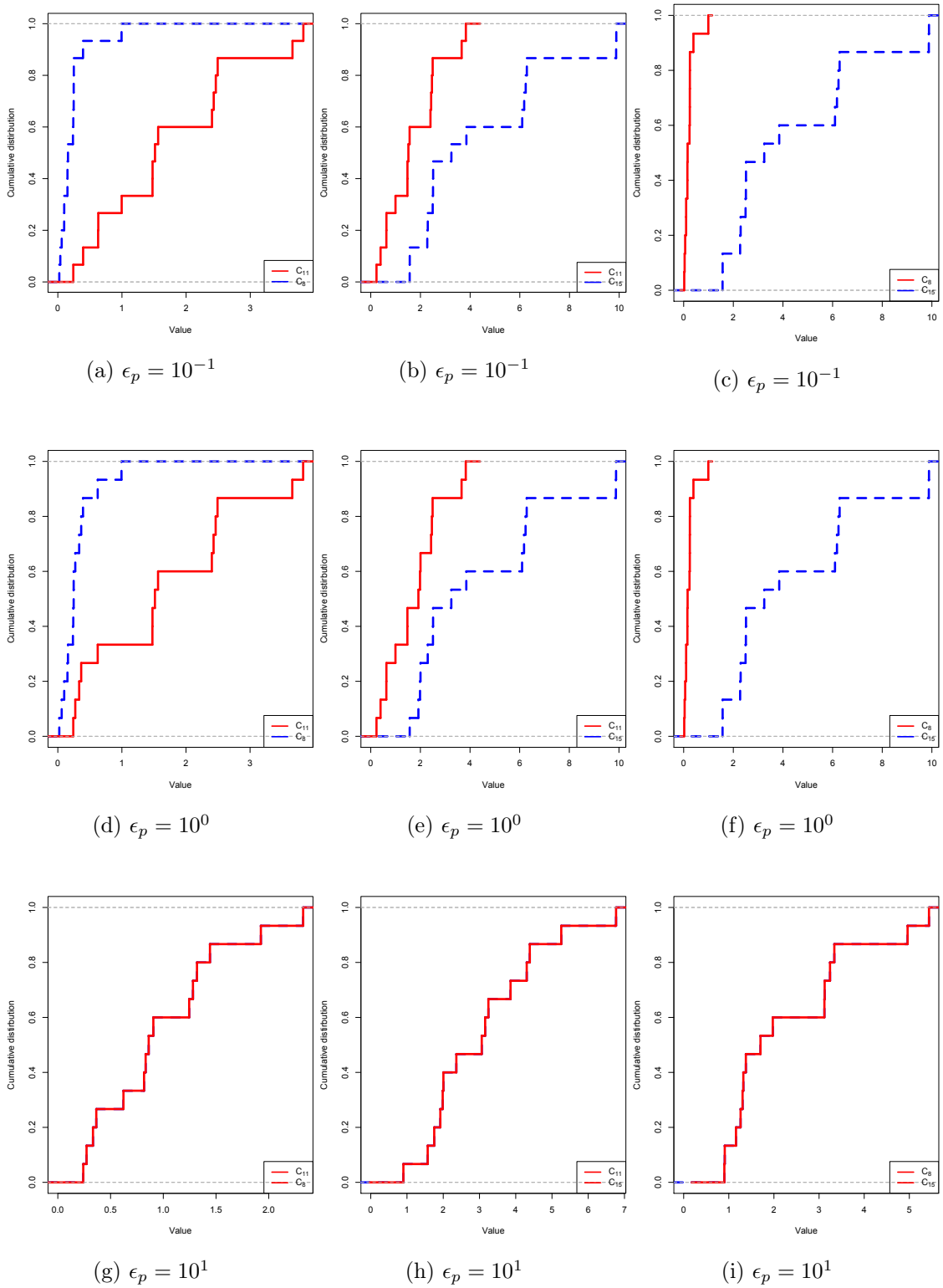
To see what happens when the order of the independent measurements changes, a permutation for problem f_{17} is used for each country. In this case, the practical significance level is $\epsilon_p = 10^0$. The obtained ranks for other problems remain the same as the ranks presented in the Table 3.18b because the order of the independent measurements is the same. For the problem f_{17}, one permutation from the independent measurements for each country is used (Table 3.19). Using the order of the independent measurements for the problem f_{17}, the ranks obtained using the *DSC* ranking scheme are 2.00, 1.00, and 3.00, respectively. Using the ranks for the other problems and the new ranks for the problem f_{17}, the *Friedman* statistic is 5.613 with a corresponding p-value 0.060. Because the p-value is greater than 0.05 the null hypothesis is not rejected meaning there is no practical difference between the countries when $\epsilon_p = 10^0$.

From this we can conclude that the order of the independent measurements influences preprocessing, which affects practical significance. In this example, only the order of independent data for one problem is changed and the result for practical significance changes. To avoid the influence from the order of the independent results, the Monte-Carlo DSC ranking scheme is proposed.

3.4.5 Multiple comparisons among three data sets by using the Monte-Carlo DSC ranking scheme

In this experiment, we use the Monte Carlo version of the DSC ranking scheme. For each country, the number of independent measurements provided for each food is 15, $n = 15$.

For each country, 15! permutations can be generated on a single problem. To avoid the influence of the order of independent measurements for each pairwise comparison, a *Monte-Carlo* simulation is performed in which 1000 combinations out of $(15!)^2$ different combinations, $N = 1000$, are randomly selected. For each combination, preprocessing and a two-sample KS test ($\alpha_T = 0.05$), are used to obtain a p-value and because 1000 p-values are obtained for each pairwise comparison but the DSC ranking scheme needs only one of them, a kernel density estimation is used to obtain the probability density function of the p-values. The mode of the probability density function is used as an appropriate p-value by the DSC ranking scheme. The results of the *Friedman* test with statistical significance, $\alpha = 0.05$ are presented in Table 3.20. These results for different practical significance levels are the same as those obtained using sequential version. The order of the data provided



(a) $\epsilon_p = 10^{-1}$ (b) $\epsilon_p = 10^{-1}$ (c) $\epsilon_p = 10^{-1}$

(d) $\epsilon_p = 10^{0}$ (e) $\epsilon_p = 10^{0}$ (f) $\epsilon_p = 10^{0}$

(g) $\epsilon_p = 10^{1}$ (h) $\epsilon_p = 10^{1}$ (i) $\epsilon_p = 10^{1}$

Fig 3.12: Cumulative distributions for f_{17}.

in these combinations does not affect the results. However, in general, the order of the independent measurements could affect the obtained results, which is seen in the previous

Table 3.19: Permutations of independent measurements obtained on the problem f_{17} for the countries C_{11}, C_8, and C_{15}.

	$\mathbf{C}_{11}$	$\mathbf{C}_8$	$\mathbf{C}_{15}$
1	3.828	0.247	6.097
2	0.628	0.247	3.245
3	3.658	0.098	2.506
4	0.394	0.022	1.569
5	0.995	0.251	6.282
6	0.630	0.150	2.295
7	0.231	0.393	2.493
8	2.463	0.150	9.877
9	1.477	0.038	2.511
10	1.564	0.235	2.274
11	2.492	0.060	3.849
12	1.476	0.993	6.175
13	1.515	0.099	9.883
14	2.403	0.249	6.237
15	2.430	0.158	1.566

example, when different order of the data for one problem changes the result.

To see what happens at the single problem level when the Monte-Carlo version is used, let us focus on combination 21 from Table 3.20, problem f_{17}, where $\epsilon_p = 10^0$. The three countries involved in this comparison are C_{11}, C_8, and C_{15}. For each pairwise comparison, 1000 permutations of the 15 independent measurements for each country on the problem f_{17} were selected. For each combination that contains one permutation from the first country and one permutation from the second country, the data is preprocessed using the Equation 3.9. This data is then used by the two-sample KS test. After processing 1000 different combinations, the ranking scheme uses a kernel density estimation to generate the p-values probability density function. The mode of the probability density function is the p-value used by the Monte Carlo version.

Figure 3.13 shows the kernel density estimation for the p-values of each pairwise comparison. The blue vertical line corresponds to the mode of the probability density function, which for (C_{11}, C_8) is 0.002, for (C_{11}, C_{15}) is 0.027, and for (C_8, C_{15}) is 0.000. The *Bonferroni correction* is then used to test the p-values and the countries are split into two disjoint sets $\{C_{11}, C_{15}\}$ and $\{C_8\}$. The obtained rankings are 2.50, 1.00, and 2.50, respectively.

This ranking scheme can be used to consider a practical significance between the compared data sets, when the data that needs to be preprocessed is coming from unpaired samples. Using it, the influence of the order of measurements obtained on the same problem can be avoided.

3.4.6 All possible pairwise comparisons

If the null hypothesis is rejected by an omnibus statistical test, we need to proceed with an appropriate post-hoc test. However, the post-hoc tests are not a subject of our approach because we introduced a ranking scheme that can be used for omnibus statistical tests, but we will used them to show some differences that exist in the case of the common approach and the DSC approach.

In our experiment, the *Friedman test* was used, and an appropriate post-hoc test when the null hypothesis is rejected will be *Nemenyi test* [58], which is used when all countries

Table 3.20: Statistical comparison of three countries using DSC.

	Countries	ϵ_p							
		10^{-9}	10^{-6}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
1	C_{11}, C_{14}, C_1	1	1	1	1	1	1	1	1
2	C_2, C_{13}, C_1	1	1	1	1	1	1	1	1
3	C_{14}, C_9, C_1	1	1	1	1	1	1	1	1
4	C_1, C_{11}, C_{13}	1	1	1	1	1	1	1	1
5	C_4, C_{13}, C_{15}	1	1	1	1	1	1	1	1
6	C_{11}, C_{14}, C_2	1	1	1	1	1	1	1	1
7	C_1, C_3, C_2	1	1	1	1	1	1	1	1
8	C_2, C_{11}, C_{15}	1	1	1	1	1	1	1	1
9	C_9, C_1, C_{15}	1	1	1	1	1	1	1	1
10	C_2, C_4, C_{15}	1	1	1	1	1	1	1	1
11	C_{11}, C_2, C_{13}	1	1	1	1	1	1	1	1
12	C_{14}, C_8, C_9	1	1	1	1	1	1	1	1
13	C_{15}, C_1, C_2	1	1	1	1	1	1	1	1
14	C_{14}, C_9, C_{11}	1	1	1	1	1	1	1	1
15	C_4, C_{13}, C_{14}	1	1	1	1	1	1	1	1
16	C_2, C_8, C_3	1	1	1	1	1	1	1	1
17	C_8, C_2, C_{13}	1	1	1	1	1	1	1	1
18	C_3, C_{15}, C_8	1	1	1	1	1	1	1	1
19	C_3, C_{14}, C_{13}	1	1	1	1	1	1	1	1
20	C_{11}, C_2, C_{13}	1	1	1	1	1	1	1	1
21	C_{11}, C_8, C_{15}	0	0	0	1	1	0	1	1
22	C_1, C_{15}, C_5	0	0	0	0	0	0	1	1
23	C_4, C_1, C_7	0	0	0	0	0	0	1	1
24	C_{14}, C_7, C_3	0	0	0	0	0	0	1	1
25	C_1, C_3, C_6	0	0	0	0	0	0	1	1

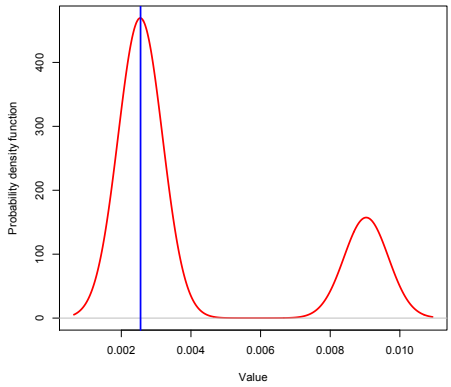
0 - indicates that the null hypothesis fails to reject, $p_{value} < 0.05$
1 - indicates that the null hypothesis fails to reject, $p_{value} \geq 0.05$
p_{value} corresponds to the p-value obtained by the *Friedman test*

are compared to each other. If we are interested in a comparison between all countries and a control country, instead of the *Nemenyi test* we can use the single-step *Bonferoni-Dunn procedure* [59], the step-down *Hommel procedure* [134], or the step-up *Holm procedure* [60]. Since we are interested to compare all countries to each other, we decided to use the *Nemenyi test*. By using it, two countries are significantly different if the corresponding average ranks differ by at least the critical difference

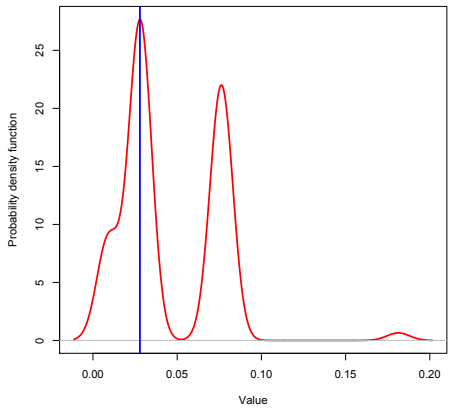
$$CD = q_\alpha \sqrt{\frac{m(m+1)}{6k}}, \tag{3.12}$$

where critical values q_α are based on the Studentized range statistic divided by $\sqrt{2}$ [12].

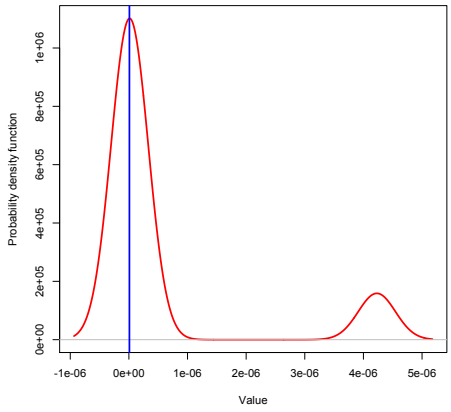
In the case when C_2, C_4, and C_{15}, are compared, the CD obtained by the *Nemenyi test* and a significance level, 0.05, is 0.71. The average ranks of the countries obtained by the common approach with averages are 2.29, 2.16, and 1.54, by using the common approach with medians are 2.43, 2.07, and 1.50, while by using the DSC approach are 1.98, 2.02, and 2.00. In the case of the common approach with averages the difference appears because the average ranks of the pair of countries, C_2 and C_{15}, differ by 0.75 that is greater than the CD used. The same happens when the common approach with medians is used, where the difference appears according to the difference between the average ranks of the pair of countries C_2 and C_{15}, that is 0.93 and is greater than 0.71. In the case of the



(a) (C_{11}, C_8)



(b) (C_{11}, C_{15})



(c) (C_8, C_{15})

Fig 3.13: Kernel density estimation for the obtained p-values for each pairwise comparison.

DSC approach the null hypothesis is not rejected and there is no need to continue with a post-hoc test, but in order to obtain the result using CD we can easily find the difference between the average ranks of the pairs of countries, from where we can see that neither of them is greater than the CD used. Also the results of the post-hoc test can be visually represented with a simple diagram such as the one presented in Figure 3.14. In this figure the results obtained by the *Nemenyi test* are presented for the common approach with averages, the common approach with medians, and the DSC approach. When comparing all countries against each other, the groups of countries that are not significantly different are connected together.

In addition if we focus only on one problem, for example f_4, we can see that the ranks obtained by the common approach using averages are 3.00, 2.00, and 1.00, respectively. The averages from the multiple measurements of each country are $4.787631e{-}09$, $3.951058e{-}09$, and $2.845056e-09$, respectively, and according to them the countries obtained different ranks because there are no ties presented. We need to note here that the countries obtained different ranks even when the difference between the averages of the pairs of countries are in some ϵ-neighborhood, in our case, $e-09$.

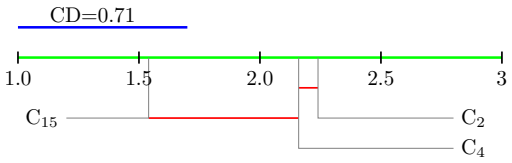
The same scenario is when the common approach with medians is used. The medians obtained are $5.25516e-09$, $3.404125e-09$, and $1.706084e-09$, respectively. According to them the countries obtained different ranks 3.00, 2.00, and 1.00, even when the differences between them are in some ϵ-neighborhood, in our case, $e-09$. This happens because there are no ties between the medians.

By using the DSC approach, the ranks obtained for f_4 are 2.00, 2.00, and 2.00, respectively. In this case, first the whole distribution of the multiple measurements between the pairs of countries are compared. The p-values obtained by the KS test are, 0.67 (C_2, C_4), 0.18 (C_2, C_{15}), and 0.18 (C_4, C_{15}). Because we have 3 pairwise comparisons and we want to control the FWER, we tested each individual hypothesis at a statistical significance level 1/3 of the desired maximum overall level, or 0.02. In our case neither of the hypotheses is rejected, so we can assume that they all have the same distribution, and the ranks obtained are the same, even when the difference between the averages are in some ϵ-neighborhood.

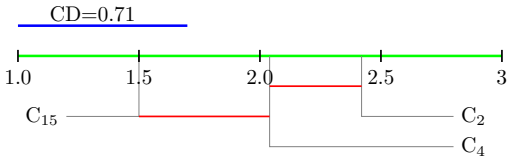
In the case when C_8, C_2, and C_{13}, are compared, the common approach with averages and the DSC approach give the same result, the null hypothesis is not rejected and there is no need to continue with a post-hoc test. The average ranks of the countries obtained by using the common approach with averages are 1.86, 2.29, and 1.84, respectively, while the average ranks obtained by using the DSC approach are 1.91, 2.07, and 2.02, respectively. By using both approaches neither of the differences between the average ranks of the pairs of countries is not greater than the CD, 0.71. The results of the post-hoc test are visually shown in Figure 3.15, and the groups of countries that are not significantly different are connected together.

In the case when C_1, C_3, and C_6, are compared, the common approach with averages and the DSC approach give the same result, the null hypothesis is rejected. The average ranks of the countries obtained by using the common approach with averages are 2.52, 2.07, and 1.41, respectively, while the average ranks obtained by using the DSC approach are 2.32, 2.23, and 1.45, respectively. In the case of the common approach with averages the difference appears according to the difference between the average ranks of C_1 and C_6, which is 1.11 and is greater than 0.71. In the case of the DSC approach the difference appears according to the difference between the average ranks of C_1 and C_6, which is 0.87 and is greater than 0.71, and the difference between the average ranks of C_3 and C_6, which is 0.78 and is greater than 0.71. The results of the post-hoc test are shown in Figure 3.16, and the groups of countries that are not significantly different are connected together.

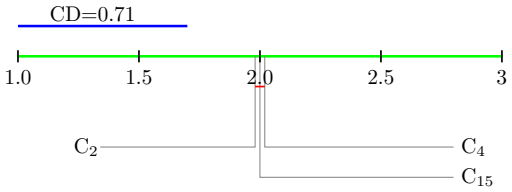
From this experiment, we can conclude that DSC also changes the results that are



(a) Comparison of all countries against each other using the *Nemenyi test* (common approach using averages). Groups of countries that are not significantly different (at $\alpha = 0.05$) are connected.

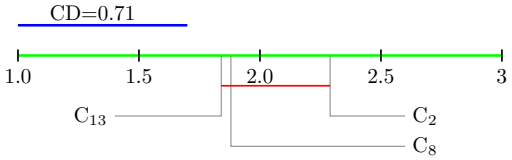


(b) Comparison of all countries against each other using the *Nemenyi test* (common approach using medians). Groups of countries that are not significantly different (at $\alpha = 0.05$) are connected.

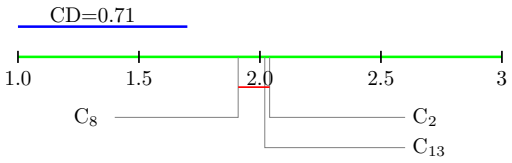


(c) Comparison of all countries against each other using the *Nemenyi test* (DSC approach). Groups of countries that are not significantly different (at $\alpha = 0.05$) are connected.

Fig 3.14: Visualization of post-hoc tests used for C_2, C_4, and C_{15}.



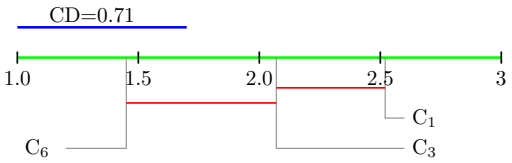
(a) Comparison of all countries against each other using the *Nemenyi test* (common approach with averages). Groups of countries that are not significantly different (at $\alpha = 0.05$) are connected.



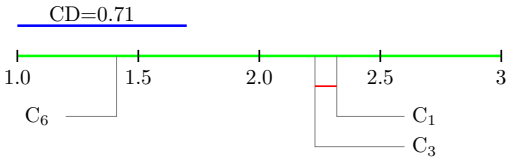
(b) Comparison of all countries against each other using the *Nemenyi test* (DSC approach). Groups of countries that are not significantly different (at $\alpha = 0.05$) are connected.

Fig 3.15: Visualization of post-hoc tests used for C_8, C_2, and C_{13}.

obtained from post-hoc procedures. This happens because post-hoc procedures are based on average ranks obtained from the omnibus statistical test, and by using the DSC ranks in a combination with omnibus statistical tests more robust results are achieved.



(a) Comparison of all countries against each other using the *Nemenyi test* (common approach with averages). Groups of countries that are not significantly different (at $\alpha = 0.05$) are connected.



(b) Comparison of all countries against each other using the *Nemenyi test* (DSC approach). Groups of countries that are not significantly different (at $\alpha = 0.05$) are connected.

Fig 3.16: Visualization of post-hoc tests used for C_1, C_3, and C_6.

3.5 Power Analysis

The power of a statistical test is defined as the probability that the test will (correctly) reject the false null hypothesis. The statistical power of the two-sample KS test for small and large samples that is used in our ranking scheme is analyzed and presented in [124]. More about the power of nonparametric statistical tests is presented through an experimental analysis in [15].

However, in this section an estimation of the power of a statistical test is made using an experimental setup by comparing countries over a number of different problems (foods). The method used is presented in [12] and is based on the idea that the criterion of what is actually false is related to the selection of the statistical test, and we can only observe the probability of the rejection of the null hypothesis, which is nevertheless related to the power. Then, we also observed the average p-value as another measure of "power" of the test: the lower the values, the more likely it is for a test to reject the null hypothesis at a set significance level. It is really important to note here that we do not change the statistical test by itself, but we propose a ranking scheme that can be used to transform the data using the distributions of the multiple measurements of each country for each problem, and after that the transformed data is to be used as input data for an appropriate omnibus statistical test. When comparing countries, samples of ten problems (foods) were randomly selected so that the probability for the problem i being chosen was proportional to $\frac{1}{1+e^{-kd_i}}$, where d_i is the difference between the ranks of the countries that are randomly chosen for that problem and k is the bias through which we can regulate the differences between the countries. Figure 3.17 represents the number of hypotheses rejected between the three countries (significance level of 0.05) and Figure 3.18 represents their associated average p-value. The dashed curve corresponds to the result obtained using the *Friedman test* with the common approach using averages, and the dot-dashed curve corresponds to the result obtained by the *Friedman test* using the DSC approach. These figures display the average p-value and the number of hypotheses rejected in comparison between C_1, C_3, and C_6, which is a comparison for which the null hypothesis is rejected by both approaches. From them, we can conclude that there is no difference between the power of the *Friedman test* by using the common approach and the DSC approach.

In Figures 3.19 and 3.20, the number of hypotheses rejected and the average p-value

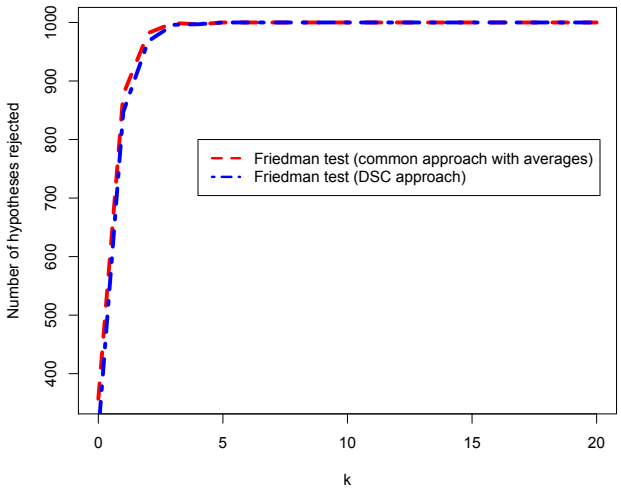


Fig 3.17: Number of hypotheses rejected between C_1, C_3, and C_6.

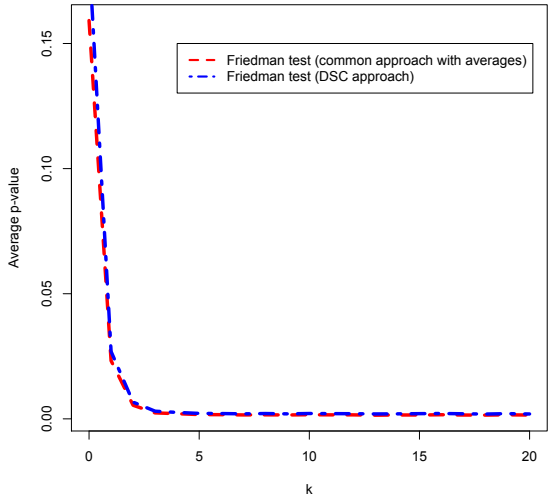


Fig 3.18: Average p-value between C_1, C_3, and C_6.

for the comparison between C_{11}, C_2, and C_{13}, are presented. For this comparison both approaches, the common approach with averages and the DSC approach, give the same result, so the null hypothesis is not rejected. From these figures we can see that both approaches give similar results. Also, we must not forget that in the case of the common approach with averages, the null hypothesis can be rejected as a result of outliers or some ϵ-neighborhood errors.

For a combination in which the common approach with averages will reject the null hypothesis, while the DSC approach fails to reject it, it is inappropriate to talk about the number of rejected hypotheses and the average p-value as measures for the power of the test. In our experimental setup we showed that in such combinations the common approach is affected by outliers or some small ϵ-neighborhood and it will always reject the

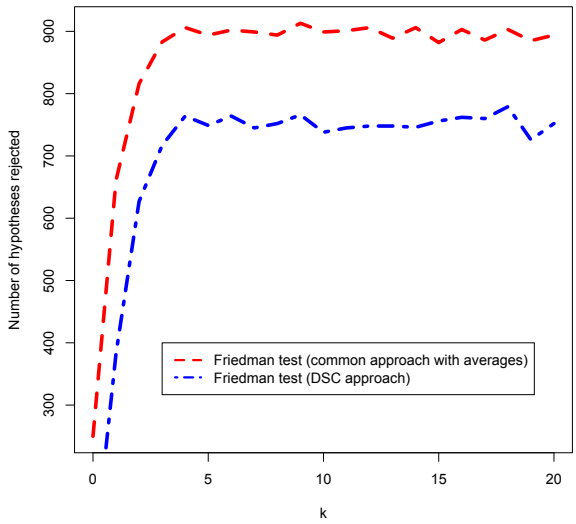


Fig 3.19: Number of hypotheses rejected between C_{11}, C_2, and C_{13}.

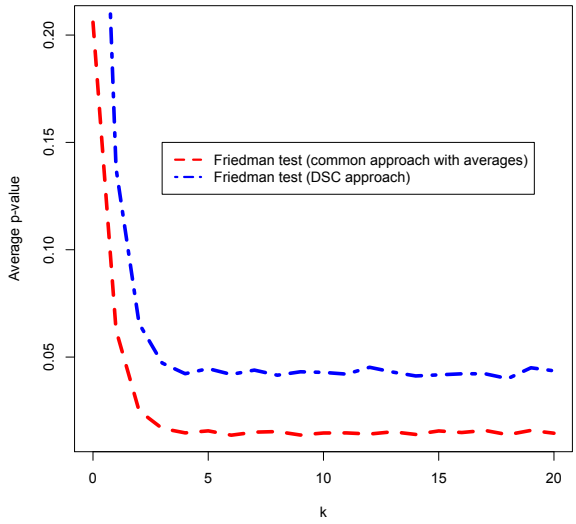


Fig 3.20: Average p-value between C_{11}, C_2, and C_{13}.

null hypothesis, while the DSC approach is more robust to these kinds of errors because it uses the whole distributions and it fails to reject the null hypothesis.

In Figures 3.21 and 3.22 an example of statistical comparison between three countries is presented according to different appropriate omnibus statistical tests using the DSC ranking scheme. Figure 3.21 represents the number of hypotheses rejected between the three countries considered with a significance level of 0.05 and Figure 3.22 represents their associated average p-value. In both of them the results are provided by using the *Fridaman test*, *Iman Davenport test*, and *Friedman aligned-rank test*, in combination with the DSC ranking scheme. The dashed curve corresponds to the result obtained from the *Friedman test*, the dot-dashed curve corresponds to the result obtained from the *Iman Davenport test*,

and the two-dashed curve corresponds to the result obtained from the *Friedman aligned-rank test*. Using them, we find that the *Friedman test*, *Iman Davenport test*, and *Friedman aligned-rank test* behave very similar.

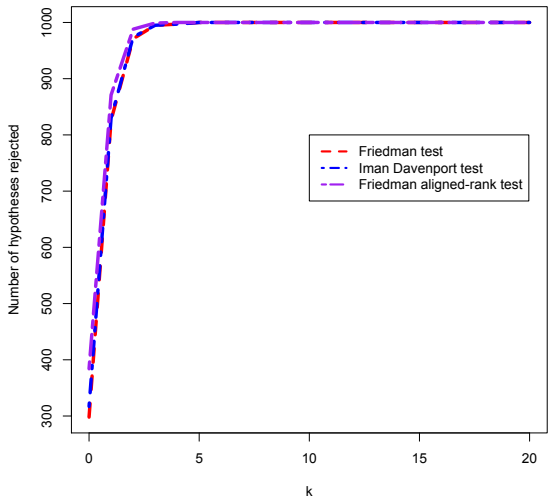


Fig 3.21: Number of hypotheses rejected between C_1, C_3, and C_6.

3.6 DSC Ranking Scheme With Different Criteria of Comparing Distributions

The first evaluation of the DSC ranking scheme was made using the two-sample *Kolmogorov-Smirnov* (KS) test. However, distributions can be compared using different criteria, statistical tests. Further, we analyze the behavior of the DSC using two different criteria, the two-sample *Kolmogorov-Smirnov* (KS) test and the *Anderson-Darling* (AD) test [27], [135].

The two-sample KS test is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare two samples. The two-sample KS test is one of the most useful and general non-parametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples (e.g., a difference with respect to location/central tendency, dispersion/variability, skewness, and kurtosis) [124].

The two-sample AD test has the same advantages mentioned for the KS test, sensibility to shape and scale of a distribution and its applicability to small samples. In addition, it has two extra advantages over the KS test. First, it is especially sensitive towards differences at the tails of distributions, and second it is better capable of detecting very small differences, even between large sample sizes. More about the comparison of the two tests when shift, scale, and symmetry of distributions are varied independently for different sample sizes can be found in [135].

To explain the difference that appears using the common approach and the two versions of the DSC, one example where the result from the common approach and both versions of the DSC differs is randomly selected and presented in detail. The explained combination is a comparison between the countries C_9, C_{14}, and C_1. For the analysis, the *Friedman test* was selected. In the case of the common approach, the null hypothesis is rejected, while

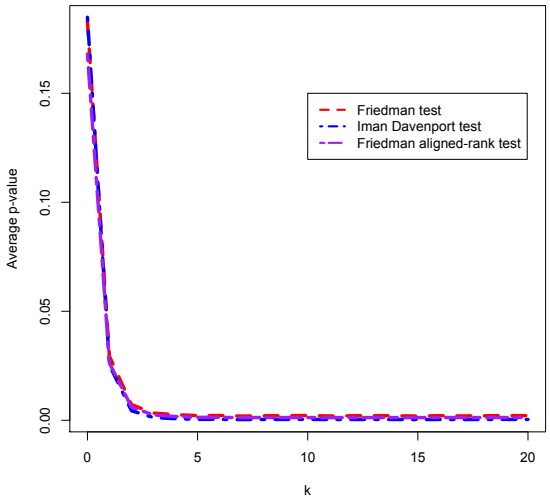


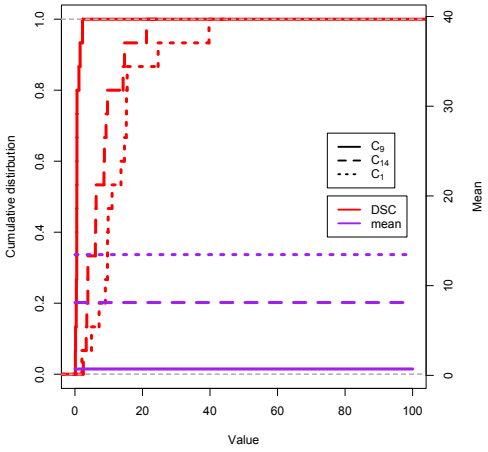
Fig 3.22: Average p-value between C_1, C_3, and C_6.

when the DSC approach is used, the null hypothesis is not rejected. The ranks obtained by the *Friedman test* using the common approach with averages and both versions of the DSC ranking scheme are presented in Table 3.21. Comparing the ranks obtained by the common approach and the DSC ranking scheme, the difference between the ranks that appears by using them can be clearly observed. To explain the difference, separate problems are discussed in detail.

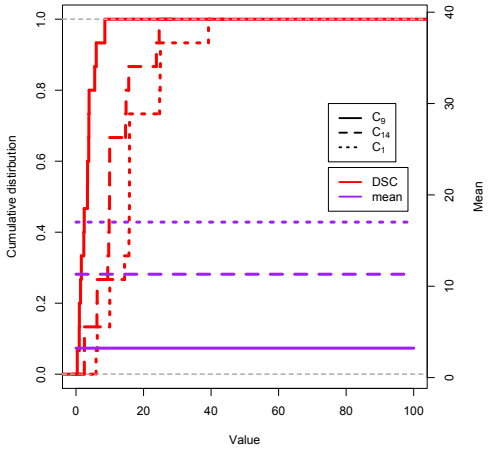
Table 3.21: Ranks for the countries C_9, C_{14}, and C_1.

F	**Common approach (*Friedman test*)**			**DSC ranking scheme (*KS test*)**			**DSC ranking scheme (*AD test*)**		
	C_9	C_{14}	C_1	C_9	C_{14}	C_1	C_9	C_{14}	C_1
f_1	3.00	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00
f_2	3.00	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00
f_3	13.00	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00
f_4	3.00	1.00	2.00	3.00	1.00	2.00	3.00	1.00	2.00
f_5	3.00	1.50	1.50	2.00	2.00	2.00	2.00	2.00	2.00
f_6	3.00	1.00	2.00	3.00	1.00	2.00	2.50	1.00	2.50
f_7	**1.00**	**2.00**	**3.00**	**1.00**	**2.50**	**2.50**	**1.00**	**2.50**	**2.50**
f_8	3.00	1.00	2.00	3.00	1.50	1.50	3.00	1.00	2.00
f_9	3.00	1.00	2.00	3.00	1.50	1.50	3.00	1.50	1.50
f_{10}	1.00	2.00	3.00	1.00	2.50	2.50	1.00	2.50	2.50
f_{11}	1.00	2.00	3.00	1.00	2.50	2.50	1.00	2.50	2.50
f_{12}	3.00	2.00	1.00	3.00	1.50	1.50	3.00	1.50	1.50
f_{13}	2.00	1.00	3.00	1.50	1.50	3.00	1.50	1.50	3.00
f_{14}	3.00	1.00	2.00	2.50	2.50	1.00	2.50	2.50	1.00
f_{15}	2.00	1.00	3.00	2.00	2.00	2.00	2.00	2.00	2.00
f_{16}	2.00	1.00	3.00	2.00	2.00	2.00	2.00	1.00	3.00
f_{17}	1.00	2.00	3.00	1.00	2.50	2.50	1.00	2.50	2.50
f_{18}	**1.00**	**2.00**	**3.00**	**1.00**	**2.50**	**2.50**	**1.00**	**2.00**	**3.00**
f_{19}	3.00	1.00	2.00	3.00	1.50	1.50	3.00	1.50	1.50
f_{20}	3.00	1.00	2.00	3.00	1.50	1.50	3.00	1.50	1.50
f_{21}	**1.00**	**2.00**	**3.00**	**2.00**	**2.00**	**2.00**	**2.00**	**2.00**	**2.00**
f_{22}	1.00	2.00	3.00	2.00	2.00	2.00	2.00	2.00	2.00

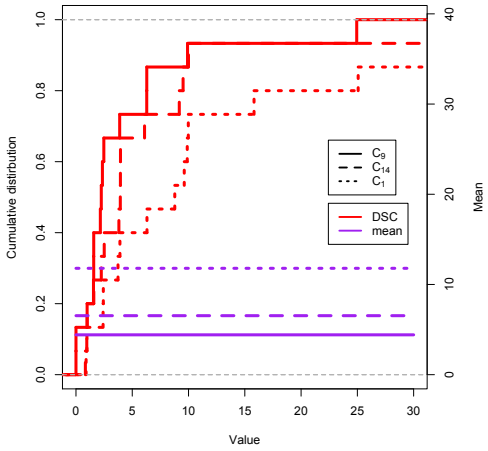
In Figure 3.23 the cumulative distributions (the step functions) and the average values



(a) f_7



(b) f_{18}



(c) f_{21}

Fig 3.23: Cumulative distributions (the step functions) and average values (the horizontal lines) for different problems of C_9, C_{14}, and C_1.

(the horizontal lines) obtained from the multiple measurements for different problems of the three countries are presented. In this figure details about the problem, f_7, are presented. The ranks obtained using the common approach with averages are 1.00, 2.00, and 3.00, and they are different because all of them have different averages. The ranks obtained using the two versions of DSC ranking scheme, by using KS test and AD test, are 1.00, 2.50, and 2.50. The DSC ranking scheme used the cumulative distributions in order to assign the ranks of the countries. From the figure, one may assume that there is no significant difference between the cumulative distributions of C_{14} and C_1, but they differ from the cumulative distribution of C_9. This result is also obtained by using the two-sample KS and AD test. The p-values obtained for the pairs of countries are 0.00 (C_9, C_{14}), 0.00 (C_9, C_1), and 0.07 (C_{14}, C_1), by using the KS test, while the p-values obtained for the same pairs of countries by using the AD test are 0.00 (C_9, C_{14}), 0.00 (C_9, C_1), and 0.02 (C_{14}, C_1). Because multiple pairwise comparisons are made, these p-values are further corrected by using the *Bonfferoni correction*. In this case, the transitivity of the matrix M_7' is satisfied,

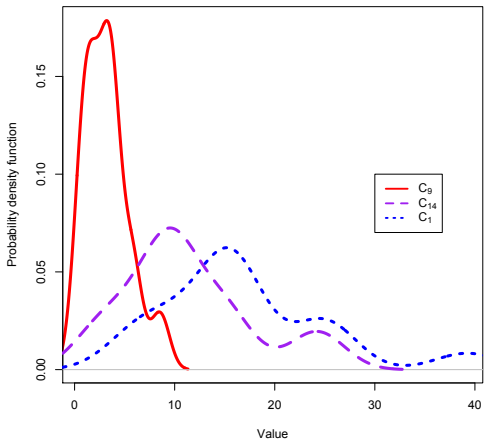


Fig 3.24: Probability density functions for f_{18} for the countries C_9, C_{14}, and C_1.

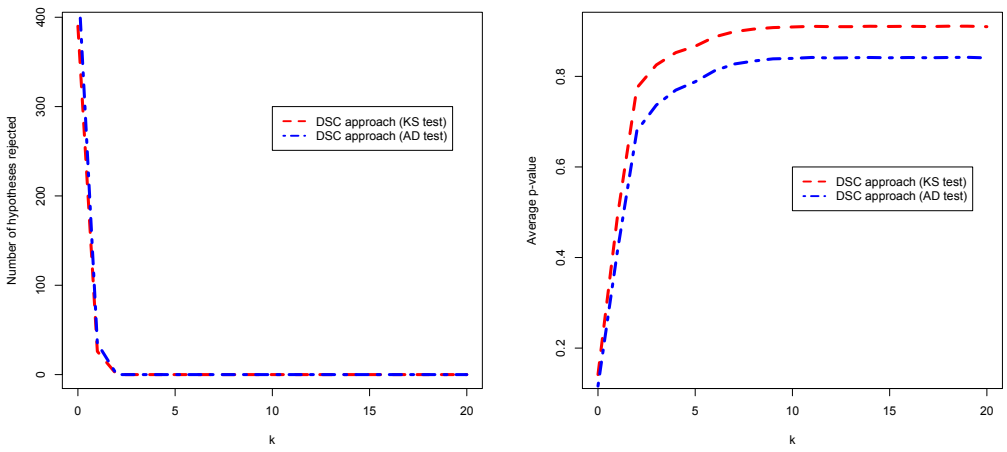
so the set of all countries is split into two disjoint sets $\{C_9\}$, and $\{C_{14}, C_1\}$, and the ranks are defined using Equations 3.4 and 3.5.

In Figure 3.23c the results for the problem, f_{21}, are presented. The ranks obtained using the common approach with averages are 1.00, 2.00, and 3.00, and they are different because all of them have a different average. The ranks obtained using the DSC ranking scheme by using the KS test and AD test they are 2.00, 2.00, and 2.00. From the figure, it is not clear if there is a significant difference between the cumulative distributions of C_9, C_{14}, and C_1. To check this, the two-sample KS test and AD test are used. The p-values obtained for the pairs of countries are 0.38 (C_9, C_{14}), 0.07 (C_9, C_1), and 0.38 (C_{14}, C_1), and 0.41 (C_9, C_{14}), 0.02 (C_9, C_1), and 0.29 (C_{14}, C_1), respectively. Because multiple pairwise comparisons are made, these p-values are further corrected using the *Bonfferoni correction*. In this case, the transitivity of the matrix $M_{21}{}'$ is satisfied, but the set of all countries is not split into disjoint sets because all countries belong to one set, $\{C_9, C_{14}, C_1\}$.

In Figure 3.23b, the results for the problem, f_{18}, are presented. This example is interesting because both versions of the DSC ranking scheme that use different criteria for comparing distributions, the KS test and AD test, give different results. For the problem f_{18}, the ranks obtained by the common approach are 1.00, 2.00, and 3.00. The ranks obtained by the DSC ranking scheme with KS test are 1.00, 2.50, and 2.50, while with the AD test are 1.00, 2.00, and 3.00. So the two different criteria used by the DSC ranking scheme give different results. The p-values obtained by using the KS test for the pairs of countries are 0.00 (C_9, C_{14}), 0.00 (C_9, C_1), and 0.03 (C_{14}, C_1). Because multiple pairwise comparisons are made, these p-values are further corrected by using the *Bonfferoni correction*. In this case, the transitivity of the matrix $M_{18}{}'$ is satisfied, so the set of all data sets is split into two disjoint sets $\{C_9\}$, and $\{C_{14}, C_1\}$. The p-values obtained by using the *AD test* for the pairs of countries are 0.00 (C_9, C_{14}), 0.00 (C_9, C_1), and 0.01 (C_{14}, C_1). Because multiple pairwise comparisons are made, these p-values are further corrected using *Bonfferoni correction*. In this case, the transitivity of the matrix $M_{18}{}'$ is not satisfied, so the countries obtain their ranks according to their averages. So, the two different criteria give different results. This result is important when we compare countries on one problem (food), while it does not influence the result when we are performing multiple-problem analysis. Even more, when we are comparing countries on one problem, it is better to use

AD test because it is more powerful and it can better detect differences than the KS test when the distributions vary in shift only, in scale only, in symmetry only, or that have the same mean and standard deviation but differ on the tail ends only [135]. Also, it requires less data than the KS test to reach sufficient statistical power. To see this difference, in Figure 3.24 the probability density functions for the data of each country obtained on the problem f_{18} are presented. From it, we can see that the KS test can not detect the small differences that exist between the probability distributions of the countries C_{14} and C_1, but the AD test can detect them.

The comparison of the statistical power of the two-sample KS test and the two-sample AD test is presented in [135]. Here we focus on power analysis between the two versions of the DSC approach that use different criteria for comparing distributions. Figure 3.25a represents the number of hypotheses rejected between the three countries considered with a significance level of 0.05 and Figure 3.25b represents their associated average p-values. From them, we can conclude that both versions of DSC approach behave similarly. The two-sample AD test has better power than the two-sample KS test [135], but this influence is not emphasized in the case of the DSC approach.



(a) Number of hypotheses rejected (b) Average p-value

Fig 3.25: Power analysis for the combination C_6, C_{14}, and C_2.

3.7 Discussion and Conclusion on DSC

Data published in some scientific papers may be unreliable because of wrong statistical analysis. This is because each statistical method requires knowledge of those conditions that must be met in order to use them. These initial conditions are not usually checked and researchers simply apply a statistical method, in most cases taken from a similar study, which is unsuitable to their data set. As a result, the conclusions from their study are incorrect. In [39], we provided a short tutorial on how to perform a statistical analysis of food data, followed by a freely available e-learning tool that automatically checks the conditions for each method and offers those methods that are appropriate for the analysis (`http://ws.ijs.si/dsc/`). The e-learning tool can help not just in interpreting results, but also in increasing awareness of using an inappropriate statistical method.

Even more, the results of the state-of-the-art methods for statistical comparison can

be affected by the presence of outliers and small ϵ-neighbourhood that exists between data values that are compared. This happens especially when one parameter is measured by different methods and needs to be compared over multiple samples, and each sample is measured several times. Our first hypothesis is that, in order to obtain more robust statistical results, the whole distribution of the data should be taken into account in the ranking process instead of using a simple statistic such as average or median, which are used to describe the data. The hypothesis is confirmed using continuous data from the field of the evolutionary computation. Experimental results show that DSC gives more robust results compared to the commonly-used approach for making statistical comparison in classical inferential statistics.

The proposed approach is based on a new ranking scheme that uses the whole distribution of the experimental data in order to rank the data sets involved in the comparison. All possible pairwise comparisons are made between the data sets and the p-values are organized in a matrix, which is further corrected using some p-value correction to control the FWER. Then, depending on the transitivity property of that matrix, the rank of each data set is obtained. The time complexity for comparing 3 data sets over 22 different samples, where each sample is measured 15 times, is 6.06 s. The same time is needed, 6.73 s, for comparing 4 data sets over 22 different samples, where each sample is measured 15 times. The multiple comparisons with a control algorithm can also be made in around 6.00 s because each pairwise comparison can be performed in parallel. The sequential DSC ranking scheme needs 6.50 s for 3 data sets over 22 different samples, where each sample is measured 15 times, while for the same parameters the Monte-Carlo DSC ranking scheme with 1000 number of evaluations needs 131.53 s. This happens because the current version of the code for Monte-Carlo DSC ranking scheme is generally programmed in a sequential way for a testing purpose. In future, this can be improved using parallel programming.

The experimental results conducted in the first three experiments (multiple comparisons among three data sets, multiple comparisons among four data sets, and multiple comparisons with a control algorithm) show that DSC approach gives more robust results compared to the commonly-used approach in case when results are affected by outliers or by a misleading ranking scheme [25]–[29]. To relate industry and research, the last two experiments (multiple comparisons among three data sets using a sequential DSC ranking scheme and a Monte-Carlo DSC ranking scheme) present more robust results about the practical significance. The approach is generally applicable in any domain including: chemical analyses, biological analyses, physical analyses, economy, evolutionary optimization, machine learning, etc.

The limitation of the method is that it currently works only with one-dimensional data and continuous data. For our future work, we are interested in extending it in order to work with high-dimensional data that can also be a subject of statistical comparison. Also, the number of data sets involved in the comparison can influence the ranking scheme that is based on multiple pairwise comparisons whose p-values are further corrected to control the FWER. The correction comes at the cost of increasing the probability of producing false negatives, or reducing the statistical power. To avoid the influence of the correction in the DSC ranking scheme, an appropriate way of comparing multiple comparisons with a control data set is to perform *Wilcoxon* test for each pairwise comparison and then obtained p-values to be combined using the equation for combing more independent p-values that are results from pairwise comparisons. Also, we will further focus to quantify the robustness of the DSC approach using artificial data that contains artificially imputed outliers. One direction is also to change it from the perspective of Bayesian statistics.

The DSC is also available as a statistical tool. It can be used via two HTTP REST API web services and a web interface developed as *Shiny* application in the *R programming*

language. The statistical tool exposes two web services through an HTTP REST API. The first web service is used to obtain ranked data by using the proposed ranking scheme, and the second one is used for statistical comparison analysis. Both services accept input parameters in JSON format and return a result in the JSON format. The first web service needs, as input parameters, data for each problem of each data set and the significance level α_{KS} that is used for the two-sample KS test in the ranking scheme. The results of the first web service are the ranked matrix, the number of data sets involved in the comparison, and one boolean variable that tells us if the required conditions for the safe use of the parametric tests are satisfied. The second web service requires, as input parameters, ranked matrix obtained from the first web service and the name of the statistical test that will be used for the statistical comparison together with the significance level for the statistical test. The results from this web service are the test statistic obtained by the statistical test that is selected, the p-value of the test, and a statement that tells us if the null hypothesis is rejected or not rejected. The web interface of the statistical API is also available. There is a template, a csv file, of how to organize the data that can be uploaded and used for statistical comparison. When we use the web interface, the first step is to validate the csv file that is uploaded. If there are errors, they are removed from the data, and the error message reports what was changed and the accepted data without errors is presented. If there are no errors reported, all data is accepted for analysis. The next step is to select the significance level, α_{KS}, for the two-sample KS test in the procedure of ranking scheme. Then, a list of appropriate statistical tests is presented and the user needs to select which statistical test will be used for the statistical comparison and also to select the value of the significance level for the selected statistical test. The results of the test are the test statistic together with the obtained p-value, and a statement that tells us if the null hypothesis is rejected or fails to reject. The web interface is developed as *Shiny* application in the *programming language R*. More information about the statistical tool can be found at the following link, $http://ws.ijs.si/dsc/$.

We proposed a novel approach for making a statistical comparison in multiple-problem scenario. The main contribution of the proposed approach is that the DSC ranking scheme is based on the whole distribution, instead of using only one statistic to describe the distribution, which can be the average or the median in the common approach. By using the common approach with averages or medians the results can be affected by the presence of the outliers and the ranking scheme that is used by some standard statistical tests when the differences between them are in some ϵ-neighborhood though the distribution is the same. By using the DSC in the cases when there is no significant statistical difference between data sets, the wrong conclusions caused by the presence of the outliers or the ranking scheme used by some standard statistical tests can be avoided.

Chapter 4

Rule-Based Named-Entity Recognition

To follow new knowledge in a particular domain that comes rapidly with new published papers or in any other electronic form, named-entity recognition (NER) methods are used to automatically identify and extract the relevant information in the domain. In this chapter, we focus on information extraction of nutrition science domain, which is an untapped domain. We provide the problem definition, followed by an overview of NERs used in biomedical domain, which is the closest domain to the nutrition science. To extract data from the dietary domain, we propose a rule-based NER. At the end, we present the evaluation results together with a discussion and the benefits of using the proposed method.

4.1 Problem Definition

Extracting domain-specific knowledge from texts is a challenging task in natural language processing (NLP). For example, definitions of domain terms/concepts are an important source of knowledge, therefore many approaches for their extraction have been proposed by NLP community [136]–[138]. Many approaches are rule-based and are based on pattern matching using mainly syntactic and lexical features, but also paralinguistic and layout information, so grammars of the domain need to be available. Studies of how to learn definitions of concepts from Slovene structured and unstructured knowledge-rich resources are presented in [136], [137]. Using rule-based approaches we can extract domain-specific concepts but if we are interested also to classify them we need to the perform another subtask of NLP. Automatic identification and classification of words or phrases that describe important concepts (entities) can be made by a process known as Named Entity Recognition (NER) [17]. From the literature review presented in Subsection 4.2, most of the NERs are corpus-based methods. Working with an untapped domain, having a corpus (the ground truth) is really a challenging task that is a time-consuming task and requires many human experts involved in the process. In such cases, an additional way to extract relevant information is to develop terminological-driven or rule-based NER. As we said before, terminological-driven NERs work by matching text phrases to concepts that exist in terminological resources. Their disadvantage is that only the concepts that exist in the resources will be recognized, but the benefit of using them is related to the frequent updates of terminological resources with new concepts and synonyms. Also, some NERs, known as rule-based, can be used. They use regular expressions that combine information from terminological resources and characteristics of the entities of interest. However, such rules also require support from human experts and time to be created. For these reasons, we propose a novel rule-based NER method for information extraction, which does

not use an annotated corpus and the rules are not associated with the entities of interest and its application to an untapped domain, which is the dietary domain, more precisely evidence-based dietary recommendations.

Several questions need to be considered to extract relevant information from evidence-based dietary recommendations. The information is related to dietary reference values, DRVs, with corresponding life stage groups, gender groups, and reference values for heights and weights for life stage and gender groups, and food composition data that usually contains information for a huge number of components, such as energy, macronutrients (e.g., protein, carbohydrate, fat), and their components (e.g., sugars, fatty acids), minerals (e.g., calcium, iron, sodium), and vitamins. The first question is how to select parts of the text (phrases) that will be candidates for the entities in which we are interested. For this purpose, a good tokenization needs to be applied, such that each phrase can be a candidate for an entity and a phrase should not contain information about more than one entity. For example, let us have the recommendation *"The recommended intake for total fiber for adults 50 years and younger is set at 38 g for men and 25 g for women, while for men and women over 50 it is 30 g and 21 g per day, respectively, due to decreased food consumption."* [139]. In this example, it is preferred to obtain *"the recommended intake for total fiber"* instead of *"fiber"* because it contains information about DRVs. Also, another question is how to extract all useful information from this recommendation and then to relate it together. The information that for *"adults 50 and younger"* the recommended intake is *"38 g for men"* and *"25 g for women"*, while the recommended intake for *"men and women over 50"* is *"30 g"* and *"21 g per day"*. Then, let us have another dietary recommendation, *"Some breakfast cereals contain 150 to 300 mg of sodium before milk is added."*. In this example, the phrase *"150 to 300 mg of sodium"* is not preferred because it contains information about more than one entity. It is preferred to have two different phrases *"150 to 300 mg"* and *"sodium"*. Then, a good and representative terminological resource(s) for each entity we are interested in needs to be selected, in order to link each phrase to find a set of candidates for each entity. Additional information is also reported in the action of the recommendation, such as *"contain"*, *"consist"*, *"should further increase"*, etc.

4.2 Related Work

Nowadays, the information extraction from the biomedical literature is a very important task in order to improve public health. Because NER methods with best performances are usually corpus-based NER methods, there is a need for annotated corpus from biomedical literature that will include the entities of interest. For this purpose, different annotated corpora are produced by shared tasks, where the main aim is to challenge and encourage research teams on natural language processing, NLP, problems.

BioNLP Shared Task 2013 [140] aims to provide a common framework for information extraction in the biomedical domain. The biological questions addressed by this task were related to the molecular biology domain and its related fields. The BioNLP Shared task 2013 consists of six tasks: gene event extraction, cancer genetics, pathway curation, corpus annotation with gene regulation ontology, gene regulation networks in bacteria, and bacteria biotopes. BioNLP Shared Task 2016 comprises three tasks that address different aspects of knowledge acquisition from text and also encompasses a wide range of biological diversity [141]–[149]. The SeeDev task [141] aimed at extracting the regulation of the seed development in plants using a rich model. The Bacteria Biotopes 3 (BB3) task [145], [147] was used for the construction of a bacteria habitat database using external ontologies. The Genia 4 (GE4) task [143] aimed at delivering a new shared task framework to construct a

knowledge base of NFkB synthesis and regulation through information extraction, IE.

BioCreative II gene mention recognition [150] was a task, where different systems were designed to identify substrings in sentences corresponding to gene name mentions. The annotated corpus was provided to the participants, on which different methods were used and results in the performances varied. The best system was a semi-supervised learning method known as alternating structure optimization (ASO) [151]. Other systems were developed by using supervised machine learning, ML, algorithms. The second best performing system used conditional random fields, CRFs, [152], the third best performing system used a combination of two support vector machines, SVMs, and one CRF [153], and the fourth best performing system used a multimodal approach with two CRFs [154]. For example, one system that is evaluated on the BioCreative II GM task training corpus is BANNER [155]. This is an open-source biomedical named-entity recognition system implemented using CRF. It represents an innovative combination of known advances beyond the existing open-source systems, in a consistent, scalable package that can easily be configured and extended with additional techniques. The work on gene mention recognition continued in BioCreative III [156], where the focus was on three tasks: cross-species gene normalization using full text, extraction of protein-protein interactions from full text, including document selection, identification of interacting proteins and identification of interacting protein pairs, and an interactive demonstration task for gene indexing and retrieval task using full text. In BioCreative IV [157], the gene ontology annotation task was reintroduced along with the following new tasks: interoperability of text mining systems, web service-based named entity recognition, and chemical/drug entity name recognition. In the chemical/drug named entity recognition two main aspects were covered, the chemical document indexing and the chemical entity mention recognition. The extraction of chemical entities from unstructured text is a very important task for different research areas, because they are related with metabolism, enzymatic reactions, potential adverse effects, etc. The systems presented are based on three general strategies: supervised ML approaches, rule/knowledge-based approaches, and chemical dictionary look-up approaches [158]. The evaluation of the systems performances was made using the CHEMDNER annotated corpus, which was provided as a part of the workshop [159]. Most systems that use supervised ML methods are based on the CRFs, some of them used SVMs, and some a combination of SVMs and CRFs. Systems were presented that use mainly rule-based methods, but these require a deep understanding of both the existing chemical nomenclature standards as well as of the CHEMDNER annotation guidelines. The use of dictionary-lookup based systems required efficient dictionary pruning and post-processing of the results. For example, becas [160] provides annotations for isolated, nested, and intersected entities. It uses dictionary-matching techniques to recognise species, anatomical concepts, microRNAs, enzymes, chemicals, drugs, diseases, metabolic pathways, cellular components, biological processes, and molecular functions. It gives an opportunity to choose the types of entities in which you are interested. It was validated against CRAFT [161], AnEM [162], and NCBI Diseases corpora [163], achieving an f-measure of 76% for genes and proteins, 95% for species, 65% for chemicals, 83% for cellular component, 92% for cells, 63% for molecular functions and biological processes, 83% for anatomical entities, and 85% for diseases. Becas[chemicals] [164] is a web application and API for recognizing and annotating chemical compounds and drugs. It is a special branch of becas API focusing on identifying of a large array of chemical substances. It uses machine learning techniques, with an optimized feature set including orthographic, morphological, natural language processing, domain knowledge, and local context features. It was validated against the BioCreative IV CHEMDNER task corpus, achieving an f-measure of 87.48% for chemical named entities. The work related to chemical named entity recognition continued also in BioCreative V

[165]–[167], where the focus was on disease and symptoms related entities and relations that exist between chemical/drug entities and disease entities. BioCreative 2016 was focused on four main tasks: applications of text mining methods in areas such as crowdsourcing, database curation, the publication process, and metagenomics; methods for annotations such as disease, phenotype, and adverse reactions in different text sources literature, clinical records and social media; methods to achieve interoperability, generalisability, and scalability in text mining: BioC [168], RDF and semantic web, among others; and the application of ontologies in text mining and text mining as an ontology builder. Over the past few years, AL algorithms are also applied on multiple biomedical tasks, such as assertion classification for clinical concepts [93], high-throughput phenotype identification using electronic health record data [94], and clinical NER [95], [96]. Recently, there are also studies that improve biomedical NER using deep learning [104], [169]–[172].

From an overview of the existing IE methods from the biomedical literature, we can see that a lot of NER methods exist in the domain of biomedical literature and they are focused on different biomedical domains. The commonly used NER methods are the corpus-based NER methods that rely on annotated corpus for the domain of interest, which is produced by the domain experts. We did not find any research that focuses on extracting dietary information from evidence-based dietary recommendations. Several studies are conducted in the dietary domain, but with different goals. For example, Xia et al. [20] presented an approach to identify rice protein resistant to *Xanthomonas oryzae pv. oryzae*, which is an approach to enhance gene prioritization by combining text mining technologies with a sequence-based approach. Co-occurrence methods were also used to identify ingredients mentioned in food labels and extracting food-chemical and food-disease relationship [18], [19], [173]. Also, we could not find an annotated corpus that can be used to developed corpus-based NER methods for dietary information.

4.3 Methodology

To extract dietary information provided in evidence-based dietary recommendations, we propose a novel rule-based NER method, called drNER [30]–[32]. It is a combination of a terminological-driven NER and rule-based NER. The difference with purely terminological-driven NERs is that we do not only use dictionaries with concepts and synonyms (as terminological resources), but we allow the reuse of some corpus-based NERs that exist for some entities. If corpus-based NERs exist for some entities we are interested in, we use them to annotate text data and then to see if some tokens have labels that correspond to entities of interest. We also combine corpus-based NERs that exist for some entities in which we are interested, following the idea of ensemble learning in order to achieve better performance than the performance obtained from any corpus-based NER alone. The difference with the rule-based NERs is that we do not use rules associated with the characteristics of the entities. This is because having rules for each of the entities we are interested in requires too much time and effort to produce them. We only used a small number of Boolean algebra rules that are not related to the characteristics of the entities, but help us define the phrases that are the entities mentions.

DrNER works with text that is composed of sentences and paragraphs. First, we split the text using sentence segmentation. Then, each sentence is split with an additional segmentation and each sentence segment is pre-processed and analyzed in two phases. The first one is detection and determination of the entities mentions. It is based on a combination of NLP methods, matrix theory, and Boolean algebra. The second one is the selection and extraction of the entities. It uses the information from the first phase by representing it using graph theory and a small set of rules that define how to extract the

useful entities. The workflow diagram of the drNER method is presented in Figure 4.1.

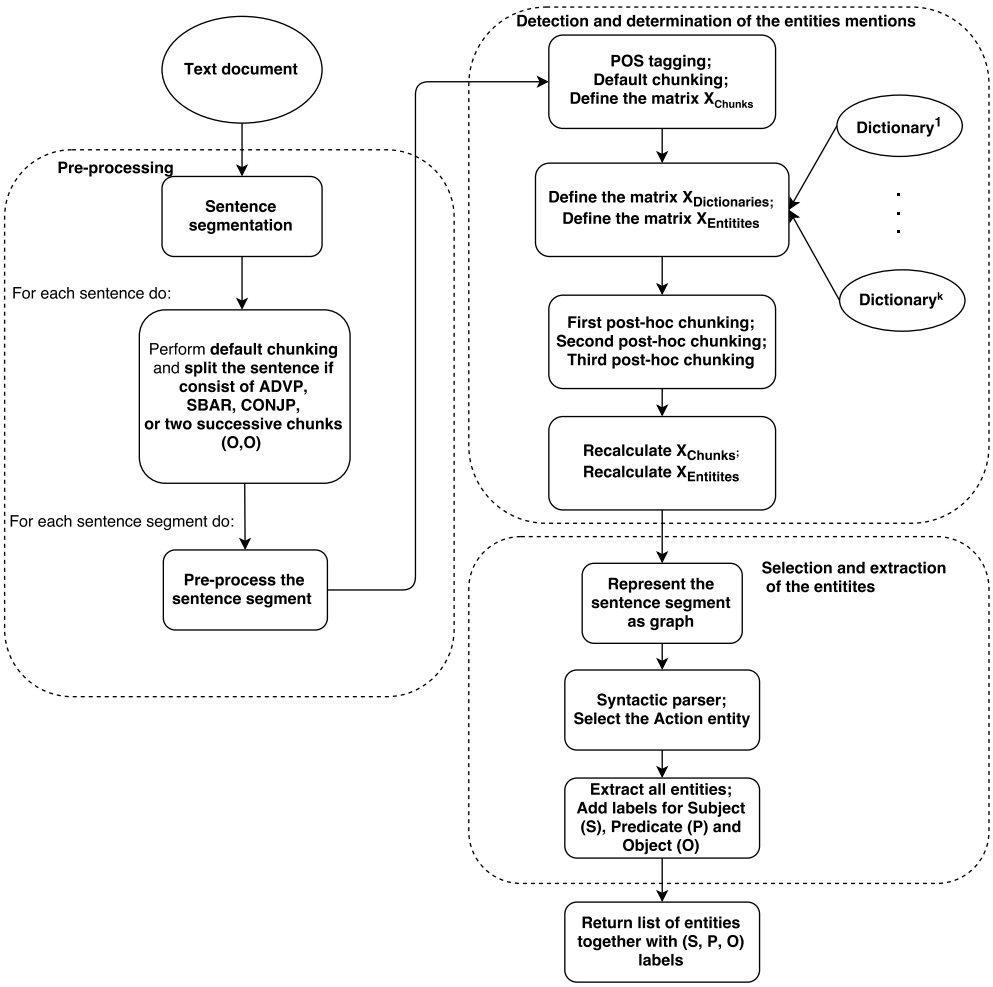


Fig 4.1: DrNER workflow diagram.

The pseudocode of the drNER is presented in Algorithm 4.1.

4.3.1 Pre-processing

Before we start with information extraction, the first step is to pre-process the text data. First, sentence segmentation is used for each text document. Then, default chunking [174] is introduced in each sentence. In our implementation, the Apache OpenNLP Maxent sentence detector is used for sentence segmentation and the Apache OpenNLP Maxent chunker is used for default chunking. They are part of the *openNLP* R package [174]. After default chunking, if the sentence consists of $ADVP$, $CONJP$, $SBAR$ chunks, or two successive chunk tokens that are (O, O), we need to split the sentence on that place or places. The $ADVP$ chunk is for an adverbial phrase, the $CONJP$ is for a conjunctive phrase, the $SBAR$ is for a subordinated clause, and (O, O) means that two successive tokens are outside of any chunk. Splitting the sentence into more segments is useful to extract more information that can stay hidden if the sentence is not split. Then for each sentence segment the double quotation marks and brackets are removed.

Algorithm 4.1: Pseudocode of drNER.

```
Apply sentence segmentation on the text document
for each sentence, Φ_i, in the set of sentences, Φ do
    Obtain Chunks vector by introducing the default chunking on Φ_i
    Split Φ_i on the position of each ADVP, CONJP, SBAR chunk, or two successive
    chunks (O,O), into set Ψ
    for each sentence segment, Ψ_i, in Ψ do
        Obtain the Tokens vector by using the word-level tokenization on Ψ_i
        Obtain the Chunks vector by introducing the default chunking on Ψ_i
        Obtain the X_Chunks matrix by using Equation 4.1
        Obtain the X_Dictionaries matrix by using Equation 4.2
        Obtain the X_Entities matrix by using Equation 4.3
        Perform the first post-hoc chunking
        Perform the second post-hoc chunking
        Perform the third post-hoc chunking
        Recalculate the X_Chunks matrix by using Equation 4.1
        Recalculate the X_Entities matrix by using Equation 4.3
        Select the Action entity by searching the predicate in the VP subtrees
        Extract all entities
        Add the labels for (Subject, Predicate, Object) using S, P, O
    end for
end for
Return list of entities together with (Subject, Predicate, Object) labels
```

4.3.2 Detection and determination of entities mentions

Let Φ be a sentence or sentence segment that contains dietary information. We start by introducing the word-level tokenization on Φ. The result is an $n \times 1$ vector, $Tokens$, whose elements are the tokens from Φ, and n is the number of tokens obtained after the tokenization. Then we continue with POS (part-of-speech) tagging and the result is an $n \times 1$ vector, POS_{Tags}, which is a collection of POS tags for Φ.

After processing the sentence at the word-level, we continue with default chunking, which segments and labels multitoken sequences called chunks. The result is an $n \times 1$ vector, $Chunks$, whose elements are chunk tokens tagged in the B-I-O tagging format.

The next step is to define an $n \times m$ matrix, X_{Chunks}, where m is the number of chunk tokens from the $Chunks$ vector that begin with the prefix B- or O. The elements of the matrix X_{Chunks} are defined by Equation 4.1, so if the token belongs to the chunk we have 1, and 0 otherwise.

$$X_{\text{Chunks}}[i,j] = \begin{cases} 1, & \text{if } Tokens[i] \in Chunks[j] \\ 0, & otherwise \end{cases}, \tag{4.1}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

Let k be the number of entities we are interested in. In order to detect and determinate the entities mentions, we try to link each chunk with the information from additional terminological resources (dictionaries) related to the domains of the entities, $Dictionary^l$, $l = 1, .., k$. Once more we would like to point out that it is not necessary that these dictionaries are standard resources, which consist of concepts with synonyms, but also they can be NER systems that exist for some entities from the domain, or even more a combination of NERs in order to achieve better performance. Then an $n \times k$ matrix,

$X_{\text{Dictionaries}}$, is defined as

$$X_{\text{Dictionaries}}[i,l] = \begin{cases} 1, & \text{if } Tokens[i] \in Dictionary^{l} \\ 0, & otherwise \end{cases}. \tag{4.2}$$

After obtaining the matrices X_{Chunks} and $X_{\text{Dictionaries}}$, an $m \times k$ matrix, X_{Entities}, is defined as

$$X_{\text{Entities}} = X_{\text{Chunks}}^{\text{T}} \cdot X_{\text{Dictionaries}}. \tag{4.3}$$

The rows of the matrix X_{Entities} correspond to the chunks and the columns are the dictionaries we included for the entities. For example if the element $X_{\text{Entities}}[i,l] \geq 1$, this means that the i-th chunk is an entity mention solution for the l-th entity. The additional step is to check if a chunk is an entity mention of more entities. If this is the case, then the chunk obtains the entity tag from the last token.

In most cases, potential entity mentions are noun phrases that are the linguistically meaningful units, but sometimes it can happen that the entity mention can consist of more noun phrases or even combinations of noun phrases with some other morphological phrases. To improve the quality of text phrases that can be entities mentions, three additional post-hoc chunkings are introduced. The first post-hoc chunking combines the information from the default chunking and the entities labels for each chunk by defining Boolean algebra rules. The second post-hoc chunking uses the information from the first post-hoc chunking and the POS tags of the tokens. The last one combines the information from the second post-hoc chunking and the entities labels for each chunk by defining rules.

- *First post-hoc chunking:* In the first post-hoc chunking, trigrams of successive chunks $(Chunk_i, Chunk_{i+1}, Chunk_{i+2})$ are analyzed and merged into one new noun chunk if the trigram is composed as $(B-NP, B-PP, B-NP)$. This is made except in cases when the two noun chunks correspond to entity mentions of different entities because merging them can lose information about one entity described by one of the noun chunks. In order not to lose this information, we define a Boolean function when this chunking needs to be performed. In Table 4.1 the Boolean function, together with the Boolean variables A and B, that in our case can be different entities, is presented. Further, a Karnaugh map Figure 4.2, also known as a K-map [175], is used to simplify the Boolean algebra expression when this chunking needs to be performed. The Boolean algebra expression or the Boolean function is obtained in

Table 4.1: The Boolean function for the first post-hoc chunking.

A	**B**	**f**
0	0	1
0	1	1
1	0	1
1	1	0

a simplified form, as a sum of minterms, as

$$f(A,B) = \neg A \vee \neg B = \neg(A \wedge B). \tag{4.4}$$

Because the number of the entities we are interested in can be greater than 2, $k > 2$, the Boolean algebra expression obtained using Equation 4.4 needs to be defined for

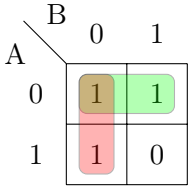


Fig 4.2: Karnaugh map of the Boolean function for the first post-hoc chunking.

each variation of pairs of entities. The number of functions is determined using the formula of the variations without repetition $V_{r,w} = \frac{r!}{(r-w)!}$, where r is the number of different elements, in our case the number of different entities, k, and w is the size of the variation or how many elements need to be selected from the set of r elements. In our case, w is 2 because we are working with a pair of entities. Then all the obtained functions are merged together with Boolean AND conditions into one expression. This expression defines whether the first chunking needs to be performed.

- *Second post-hoc chunking:* In the second post-hoc chunking, trigrams of successive chunks obtained by the first post-hoc chunking $(Chunk_i, Chunk_{i+1}, Chunk_{i+2})$ are analyzed and merged into one new noun chunk if the trigram is composed as $(B-NP, B-VP, B-NP)$ and the first noun chunk has a POS tag that is a Wh-pronoun [73], such as who, what, which, etc.

- *Third post-hoc chunking:* In the third post-hoc chunking, bigrams of successive chunks obtained by the second post-hoc chunking $(Chunk_i, Chunk_{i+1})$ are merged into one new noun chunk if the bigram is composed as $(B-NP, B-NP)$ and only one of the noun chunks is labeled as an entity of interest, or both of them have the same label.

After performing the three post-hoc chunkings, the matrix X_{Chunks} needs to be recalculated because the number of chunks is different from the number obtained by the default chunking. At the end of the first phase, the matrix X_{Entities} is recalculated and their columns correspond to the sets of entity mentions for each entity.

4.3.3 Selection and extraction of the entities

The result from the first phase of the drNER method are sets of entities mentions for each entity. The next step is the second phase in which the entities mentions form the sets that contribute to the dietary information need to be selected.

For this purpose, the sentence or sentence segment is represented as a graph, in which each chunk is connected only with its neighbors (predecessor and successor chunk). Then the start or initial node of the graph from where the search for all entities begins, is selected using syntactic bracketing or tree parsing. Each sentence or sentence segment, Φ, is represented by the parser as a tree having three children: a noun phrase (NP), a verbal phrase (VP) and a full stop (.). In addition, each sentence is formed as a combination of three parts, *Subject*, *Predicate*, and *Object* [176]. The *Subject* is the person or a thing who or which carries out the action of the verb. The *Predicate* in a sentence is what tells us about what a person or a thing does or did, or what happened to a person or to a thing. The *Object* is the person or a thing upon whom or upon which the action of the verb is carried out.

The initial node of the graph is the predicate of the sentence. The search for the predicate is performed in VP. The initial node can be found in the following subtrees

VB (verb, base form), VBD (verb, past tense), VBG (verb, present participle or gerund), VBN (verb, past participle), VBP (verb, present tense, not 3rd person singular), VBZ (verb, present tense, 3rd person singular), and MD (verb, modal). Further, from all solutions returned by searching for the predicate, the initial node is the verb chunk that is closest to the root of the sentence (number of edges from the verb node to the root node) and it is located in the verbal phrase that is closest to the root. The extracted predicate is stored in an entity called *Action*. We need to note here that because of the segmentation defined in the pre-processing step, which is done in order to extract all relevant information, it can happen that some sentence segments do not have a VP, so the *Action* entity is not returned.

If the *Action* entity is selected, all other entities of interest need to be selected. Because it can happen that no entity mention is the subject in the sentence, one additional entity called *Group* is added, into which the noun chunks that perform the action are stored. The *Group* entity is searched in the predecessor chunks from the *Action* entity that is selected. The search starts from the *Action* entity and then goes back to the beginning of the sentence. The results are the successive noun chunks that can also be separated by punctuation.

In order to know on which side of the *Action* entity the extracted entities are located, one of the labels S, P, or O that indicate $(Subject, Predicate, Object)$ is added to each extracted entity. The *Action* entity has the label P because it is the predicate of the sentence. All entities that are predecessor chunks of the *Action* entity have the label S as they are subjects in the sentence. The entities that are successor chunks of the *Action* entity have the label O because they correspond to the objects in the sentence.

Two scenarios of entities selection exist. In the first one if the *Action* entity is not selected, then all the entities mentions from the X_{Entities} matrix are extracted. In the second scenario, only one *Action* entity is returned. Then for each entity using the set of its entities mentions, the entity mention or the chunk that is closest to the *Action* entity is selected, according to the number of edges between the candidate and the *Action* entity in the graph. If the set of entities mentions consists of more candidates for the same entity, they are extracted if they are on the same side from the *Action* entity as the one extracted or they are on the other side of the *Action* entity, and there is no additional verbal phrase in this part of the sentence.

4.4 Experimental Results

In this section, we present the result of evaluation of the proposed drNER method in the domain of evidence-based dietary recommendations. The main basic entities in the domain are the *Food* entity, *Nutrient* entity, and *Quantity/Unit* entity. For a better understanding of how the method works, we start by explaining the dictionaries or the terminological resources used for the drNER, followed by examples of pre-processing, first, second and third post-hoc chinking, and how the method works on a single sentence. More information about the rules applied for the three post-hoc chunkings is presented in Appendix C. Further, the construction of the heterogeneous test corpus that consists of text paragraphs is explained. Then, the obtained result using the explained test corpus is presented. Finally, we compare the methodology of the drNER with the methodologies of some other approaches that exist in biomedical domains.

4.4.1 Dictionaries

For the *Quantity/Unit* entity, an ontology, called Units of Measurements Ontology (UO) [177]–[179], is used. The UO is currently being used in many scientific resources for the standardized description of measurement units. From it, the names of the units together with their symbols are extracted. In addition, a list of measurement units that are used for recipes, such as tablespoon, teaspoon, etc. is added.

For the *Nutrient* and the *Food* entity, dictionaries are constructed using the outputs of different NERs appropriate for the entity.

For the *Nutrient* entity, a combination of becas API [160], becas[chemicals] API [164] and a semantic tagger, known as USAS online English semantic tagger [180]–[182], is used. Both, becas and becas[chemicals], are web-services-based and corpus-based NER developed as a prat of BioCreative IV CHEMDNER task. Becas provides annotations for isolated, nested, and intersected entities. It uses dictionary-matching techniques to recognise species, anatomical concepts, microRNAs, enzymes, chemicals, drugs, diseases, metabolic pathways, cellular components, biological processes, and molecular functions. It also gives an opportunity to choose the types of entities. Becas[chemicals] is a web application and API for recognizing and annotating of chemical compounds and drugs. It is a special branch of becas API focused on the identification of a large array of chemical substances. It uses machine-learning techniques, with an optimized feature set including orthographic, morphological, natural language processing, domain knowledge, and local context features. The USAS online English semantic tagger is a part of the UCREL semantic analysis system, which is a framework of automatic semantic analysis of text that has been designed across a number of research projects since 1990. The USAS version contains 21 major entity labels, with the possibility of subdivision for some of them. For our purpose, the focus is on one category with the semantic label O, which is for terms related to substances, materials, objects, and equipment. From this entity label, we use only one subcategory O_1 that is for terms relating to substances and materials generally. The idea of using a combination of NER systems comes from the idea of ensemble learning in order to achieve a better performance of identifying the chemical entities that could be obtained from any of them alone. So if a token is annotated by at least one of these systems, as a chemical entity using becas and becas[chemicals] or O_1 using the USAS tagger, then we can assume that the token belongs to this dictionary.

For the *Food* entity the same semantic tagger as the *Nutrient* entity, known as USAS online English semantic tagger, is used. The focus is on two categories. The first one is the category for terms related to Food and Farming, F. From it, four subcategories are used. The first subcategory is for terms related to food and food preparation, F_1, the second is for terms related to drinks and drinking, F_2, the third is for terms related to cigarettes and drugs, F_3, and the fourth for terms related to agriculture and horticulture, F_4. The second category is for terms related to Life and Living things, L. From it, two subcategories are used. The first one is for terms related to living creatures (e.g., non-human), L_2, and the second subcategory is for terms related to plants and plant-life, L_3. In Table 4.2, the summary of the USAS English tagger categories and subcategories is presented. So if a token is annotated by this system as F_1, F_2, F_3, F_4, L_2, or L_3, then we can assume that the token belongs to this dictionary.

It is not possible to provide information about the size of various dictionaries being used because they are not classical dictionaries that consist of concepts with synonyms. For the *Quantity/Unit* entity, an ontology is used together with kitchen-related units. For the *Nutrient* entity, three corpus-based NER systems are used, so the results provided as annotations from these three systems are used and combined. Also, for the *Food* entity, a corpus-based NER system is used, so the results provided as annotations are used.

In Table 4.3, the summary of the dictionaries used for each entity is presented.

Table 4.2: USAS categories.

Entity	**USAS category**	**USAS subcategory**
Nutrient entity	**O** - Substances, materials, objects, and equipment	$\mathbf{O}_1$ - Substances and materials
Food entity	**F** - Food and farming	$\mathbf{F}_1$ - Food and food preparation
		$\mathbf{F}_2$ - Drinks and drinking
		$\mathbf{F}_3$ - Cigarettes and drugs
		$\mathbf{F}_4$ - Agriculture and horticulture
	L - Life and living things	$\mathbf{L}_2$ - Living creatures (e.g., non-human)
		$\mathbf{L}_3$ - Plants and plant life

Table 4.3: Dictionaries.

Entity	**Dictionary**
Quantity/Unit entity	UO ontology
	Recipe units
Nutrient entity	becas API
	becas[chemicals] API
	USAS English semantic tagger
Food entity	USAS English semantic tagger

4.4.2 Examples

We present four examples to show how different parts of the proposed method work. In the first one, an example of the pre-processing is presented to explain the difference between sentence and sentence segments. The second one includes examples for the first, second, and third post-hoc chunkings. In the third example, we present how the proposed method works on a single sentence by providing details for both phases. The forth example provides the results obtained for 15 sentences and helps the readers get more familiar with the concepts from the dietary domain that need to be extracted.

Example of pre-processing

To explain the difference between sentence and sentence segments, we continue with an example. Let one sentence, obtained from sentence segmentation of a text document, be *"The recommended intake for total fiber for adults 50 years and younger is set at 38 g for men and 25 g for women, while for men and women over 50 it is 30 g and 21 g per day, respectively, due to decreased food consumptions."* [139]. The result of the default chunking on this sentence is presented in Table 4.4. The column *Tokens* corresponds to the tokens obtained by the word-level tokenization and the column *Chunk tokens* corresponds to the

chunk token obtained by the default chunking. Further, because this sentence consists of two $ADVP$ chunks, it needs to be split in that places. After splitting, the obtained sentence segments are: *"The recommended intake for total fiber for adults 50 years and younger is set at 38 g for men and 25 g for women"*, *"For men and women over 50 it is 30 g and 21 g per day"*, and *"Due to decreased food consumptions"*. These sentence segments are further used by the proposed method.

Table 4.4: Default chunking result for one sentence.

Token id	**Tokens**	**Chunk tokens**	**Token id**	**Tokens**	**Chunk tokens**
1	The	B-NP	26	while	**B-ADVP**
2	recommended	I-NP	27	for	B-PP
3	intake	I-NP	28	men	B-NP
4	for	B-PP	29	and	I-NP
5	total	B-NP	30	women	I-NP
6	fiber	I-NP	31	over	B-PP
7	for	B-PP	32	50	B-NP
8	adults	B-NP	33	it	B-NP
9	50	B-NP	34	is	B-VP
10	years	I-NP	35	30	B-NP
11	and	O	36	g	I-NP
12	younger	B-NP	37	and	O
13	is	B-VP	38	21	B-NP
14	set	I-VP	39	g	I-NP
15	at	B-PP	40	per	B-PP
16	38	B-NP	41	day	B-NP
17	g	I-NP	42	,	O
18	for	B-PP	43	respectively	**B-ADVP**
19	men	B-NP	44	,	O
20	and	O	45	due	B-PP
21	25	B-NP	46	to	I-PP
22	g	I-NP	47	decreased	B-NP
23	for	B-PP	48	food	I-NP
24	women	B-NP	49	consumptions	I-NP
25	,	O	50	.	O

Example of the first, second and third post-hoc chunking

To see how the first post-hoc chunking works, let us focus on one example using the sentence *"People of any age who are African Americans should further reduce sodium intake to 300 mg per day."* [139]. This sentence does not consist of $ADVP$, $CONJP$, $SBAR$ chunks, or two successive chunks that are (O, O), so it is not split into sentence segments. The results of the default chunking and linking each token to a dictionary is presented in Table 4.5. The *Tokens* column corresponds to the result of the word-level tokenization. The *POS tags* column corresponds to the result of the POS tagging. The *Chunk tokens* column corresponds to the result of the default chunking, where each chunk token is presented in B-I-O tagging format and the beginning of each new chunk is marked with the symbol *. The *Food*, *Nutrient*, and *Quantity/Unit* columns correspond to the linking of each token to the dictionaries used for each entity. In the column $Chunk_1$ *tokens* the result of the first post-hoc chunking is presented, here only the new chunks formed are presented in bold font. From this table, we can see how each chunk consists of one or more chunk tokens. For example, *"sodium intake"* is one noun chunk that consists of two chunk tokens ($B-NP, I-NP$). To perform first post-hoc chunking we need to calculate matrix X_{Entities}, which is presented in Table 4.6.

Table 4.5: Example of the first post-hoc chunking.

Tokens	**POS tags**	**Chunk tokens**	**Food**	**Nutrient**	**Quantity/Unit**	**Chunk$_1$ tokens**
People	NNS	B-NP *	0	0	0	**B-NP** *
of	IN	B-PP *	0	0	0	**I-NP**
any	DT	B-NP *	0	0	0	**I-NP**
age	NN	I-NP	0	0	0	**I-NP**
who	WP	B-NP *	0	0	0	B-NP *
are	VBP	B-VP *	0	0	0	B-VP *
African	JJ	B-NP *	0	0	0	B-NP *
Americans	NNP	I-NP	0	0	0	I-NP
should	MD	B-VP *	0	0	0	B-VP *
further	RBR	I-VP	0	0	0	I-VP
reduce	VB	I-VP	0	0	0	I-VP
sodium	NN	B-NP *	0	**1**	0	B-NP *
intake	NN	I-NP	0	0	0	I-NP
to	TO	B-PP *	0	0	0	B-PP *
300	CD	B-NP *	0	0	0	**B-NP** *
mg	NN	I-NP	0	0	**1**	**I-NP**
per	IN	B-PP *	0	0	0	**I-NP**
day	NN	B-NP *	0	0	0	**I-NP**
.	.	O	0	0	0	O

[1] indicates the result of the first post-hoc chunking
* indicates a beginning of new chunk

Table 4.6: Matrix that contains candidates for each entity.

Chunks	**Chunk tokens**	**Food**	**Nutrient**	**Quantity/Unit**
People	B-NP *	0	0	0
of	B-PP *	0	0	0
any age	B-NP *	0	0	0
who	B-NP *	0	0	0
are	B-VP *	0	0	0
African Americans	B-NP *	0	0	0
should further reduce	B-VP *	0	0	0
sodium intake	B-NP *	0	**1**	0
to	B-PP *	0	0	0
300 mg	B-NP *	0	0	**1**
per	B-PP *	0	0	0
day	B-NP *	0	0	0
.	O	0	0	0

* indicates a beginning of new chunk

The information from the matrix X_{Entities} is used for the first post-hoc chunking. In order not to lose information by applying it, the Boolean function must be defined. In our case, we are interested in three entities, *Food*, *Nutrient*, and *Quantity/Unit*. By using Equation 4.4, the function is defined for all variations of pairs of entities. In our case, the number of variations is $V_{3,2} = 6$. Let the trigram of successive chunks be $(Chunk_i, Chunk_{i+1}, Chunk_{i+2})$. Then, the Boolean function is defined as

$$\begin{aligned}&\neg(Nutrient_i \wedge Quantity/Unit_{i+2}) \wedge \neg(Quantity/Unit_i \wedge Nutrient_{i+2})\\ &\wedge \neg(Nutrient_i \wedge Food_{i+2}) \wedge \neg(Food_i \wedge Nutrient_{i+2})\\ &\wedge \neg(Food_i \wedge Quantity/Unit_{i+2}) \wedge \neg(Quantity/Unit_i \wedge Food_{i+2}),\end{aligned} \tag{4.5}$$

which is true only if there is no pair for which both values are one. In the example, there are three trigrams of successive chunks that satisfy the condition $(B-NP, B-PP, B-NP)$. They are (*"People"*, *"of"*, *"any age"*), (*"sodium intake"*, *"to"*, *"300 mg"*), and (*"300 mg"*, *"per"*, *"day"*). For the first one, (*"People"*, *"of"*, *"any age"*), the Boolean function defined by Equation 4.5 is true because all entity labels are 0. So this post-hoc chunking needs to be performed and the chunks are merged into one new noun chunk, *"People of any age"*. For the trigram (*"sodium intake"*, *"to"*, *"300 mg"*), the Boolean function is false because the label $Nutrient_i = 1$ and the label $Quantity/Unit_{i+2} = 1$, so there is a pair where both values are one. According to this, $\neg(Nutrient_i \wedge Quantity/Unit_{i+2}) = 0$ and the value of the Boolean function is false, so in this case the first post-hoc chunking should not be performed. In this example, if we merged the chunks we will lose information about one entity. For the last trigram (*"300 mg"*, *"per"*, *"day"*), the label $Quantity/Unit_i = 1$, while other labels are $Nutrient_{i+2} = 0$, $Food_{i+2} = 0$, so the Boolean function is true. In this case, the first post-hoc chunking should be performed and the new chunk is *"300 mg per day"*.

To see how the second post-hoc chunking works, the obtained result from the example of the first post-hoc chunking is used. In this post-hoc chunking we are interested in the trigrams of successive chunks that satisfy the condition $(B-NP, B-VP, B-NP)$. By using the obtained result from the first post-hoc chunking, which is presented in Table 4.5, column *Chunk*$_1$ *tokens*, there are two trigrams that satisfied this condition, (*"who"*,*"are"*,*"African Americans"*) and (*"African Americans"*,*"should further reduce"*,*"sodium intake"*). From them, only the first one has the first noun chunk that is a Wh-pronoun, so they are merged together, *"who are African Americans"*.

By using the obtained result from the example of the second post-hoc chunking, there are no bigrams of chunks that satisfied the condition $(B-NP, B-NP)$, and only one of the noun chunks is labeled as an entity of interest, or both of them have the same label.

The results of post-hoc chunkings are shown in Table 4.6, where columns *Chunk*$_1$ *tokens*, *Chunk*$_2$ *tokens*, and *Chunk*$_3$ *tokens* present the results of the first, second, and third post-hoc chunking respectively.

Example of drNER performed on a single sentence

To demonstrate how drNER works, we give an example where the focus is only on a single sentence that provides dietary information, and let us consider the same recommendation given in the example of the first, second and third post-hoc chunking. Let Φ_1 "*People of any age who are African Americans should further reduce sodium intake to 300 mg per day.*" [139]. Table 4.7 gives results from the first phase of the NER method for Φ_1. In the column *Chunk*$_1$ *tokens* the result of the first post-hoc chunking is presented, where new chunks formed by this chunking are presented in bold font. The new chunks formed by the second post-hoc chunking are presented in bold font in the *Chunk*$_2$ *tokens* column. In the

Table 4.7: The first phase of the drNER method for Φ_1.

Tokens	**POS tags**	**Chunk tokens**	**Food**	**Nutrient**	**Quantity/Unit**	**Chunk$_1$ tokens**	**Chunk$_2$ tokens**	**Chunk$_3$ tokens**
People	NNS	B-NP *	0	0	0	**B-NP** *	B-NP *	B-NP *
of	IN	B-PP *	0	0	0	**I-NP**	I-NP	I-NP
any	DT	B-NP *	0	0	0	**I-NP**	I-NP	I-NP
age	NN	I-NP	0	0	0	**I-NP**	I-NP	I-NP
who	WP	B-NP *	0	0	0	B-NP *	**B-NP** *	B-NP *
are	VBP	B-VP *	0	0	0	B-VP *	**I-NP**	I-NP
African	JJ	B-NP *	0	0	0	B-NP *	**I-NP**	I-NP
Americans	NNP	I-NP	0	0	0	I-NP	**I-NP**	I-NP
should	MD	B-VP *	0	0	0	B-VP *	B-VP *	B-VP *
further	RBR	I-VP	0	0	0	I-VP	I-VP	I-VP
reduce	VB	I-VP	0	0	0	I-VP	I-VP	I-VP
sodium	NN	B-NP *	0	**1**	0	B-NP *	B-NP *	B-NP *
intake	NN	I-NP	0	0	0	I-NP	I-NP	I-NP
to	TO	B-PP *	0	0	0	B-PP *	B-PP *	B-PP *
300	CD	B-NP *	0	0	0	**B-NP** *	B-NP*	B-NP *
mg	NN	I-NP	0	0	**1**	**I-NP**	I-NP	I-NP
per	IN	B-PP *	0	0	0	**I-NP**	I-NP	I-NP
day	NN	B-NP *	0	0	0	**I-NP**	I-NP	I-NP
.	.	O	0	0	0	O	O	O

[1] indicates the result of the first post-hoc chunking
[2] indicates the result of the second post-hoc chunking
[3] indicates the result of the third post-hoc chunking
* indicates a beginning of new chunk

Chunk$_3$ tokens column the result of the third post-hoc chunking is presented and in this example nothing is changed by applying this chunking.

First, by using Equation 4.3, the X_{Entities} matrix is recalculated using *Chunk$_3$ tokens*, where rows correspond to the different chunks and columns correspond to the entities, $Food$, $Nutrient$, and $Quantity/Unit$. The X_{Entities} matrix has 6 rows (different chunks) and 3 columns. The $Food$ column gives the set of the entities mentions for the $Food$ entity, which in our example is an empty set because the dietary recommendation does not consist of food entities. The $Nutrient$ column gives the set of entities mentions for the $Nutrient$ entity, and is a set with one element that is "sodium intake" identified by the row and the chunk that has a nonzero element in the $Nutrient$ column. The $Quantity/Unit$ column gives the set of the entities mentions for the $Quantity/Unit$ entity, and is a set with one element that is "300 mg per day".

After the first phase, the recommendation Φ_1 is represented as undirected graph, where each chunk is connected only with its neighbors. In Figure 4.3 a graphic representation of the recommendation Φ_1 is presented.

The first step of the second phase is to select the initial node of the graph or the *Action* entity from where the search for all entities will start. To select the *Action* entity the parse tree of the recommendation Φ_1 is used. In Figure 4.4 we present the parse tree of the recommendation Φ_1, from which we search for the predicate in the verbal phrases. The result is the verb "should" from the MD subtree, since it is closest to the root of the sentence. So, the chunk that consists of the returned verb, "should further reduce", is selected as the *Action* entity.

The last step of the second phase is to select all other important entities. By using the second scenario (since the *Action* entity is returned), we found one $Nutrient$ entity

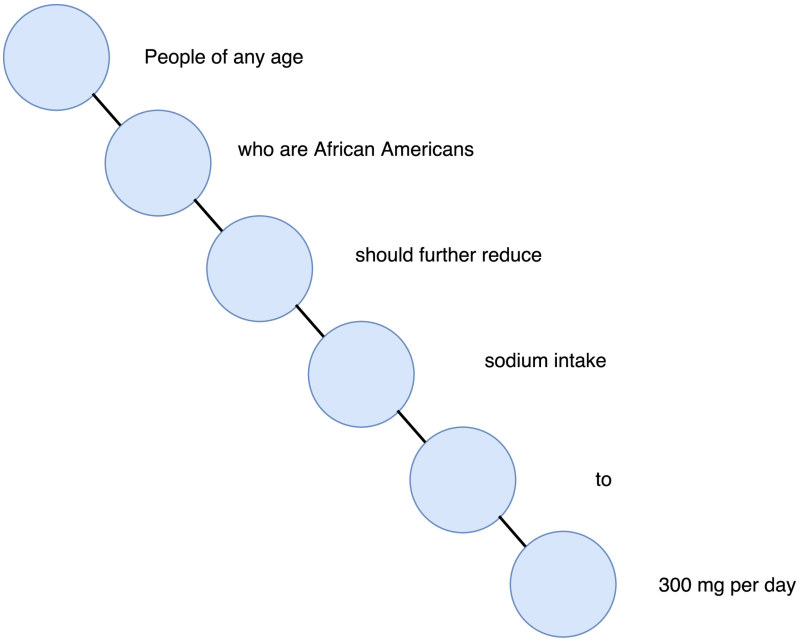


Fig 4.3: Graphic representation of the recommendation Φ_1.

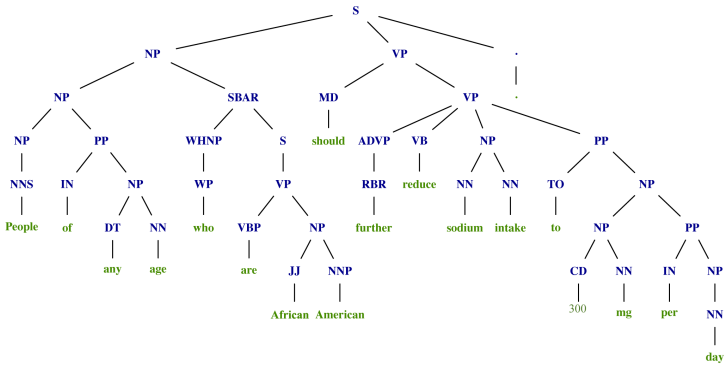


Fig 4.4: Parse tree of Φ_1.

"sodium intake", one *Quantity/Unit* entity "300 mg per day", and for the *Group* entity we obtained "People of any age" and "who are African Americans", while we did not find *Food* entity because there are no food related terms in the recommendation. At the end, the labels for the *Subject*, *Predicate* and *Object* are added ("People of any age", S_1), ("who are African Americans", S_1), ("should further reduce", P_1), ("sodium intake", O_1), and ("300 mg per day", O_1), which is also the result from our method. The index of the labels indicates from which part of the sentence the entity is extracted. In this example it is 1, because the recommendation does not contain any $ADVP$, $CONJP$, $SBAR$, or two successive chunks that are (O,O), so it is not split at the beginning.

Example of results obtained by drNER from multiple sentences

In this example we present the results obtained for 13 sentences that are extracted from the Food and Nutrition Information Centre of United States Department of Agriculture [139]. This example helps the readers to get more familiar with the concepts from the dietary domain that need to be extracted. The results are presented in Table 4.8.

Table 4.8: Information extraction of 13 dietary recommendations

	Recommendation	Group	Action	Food	Nutrient	Quantity/Unit
1.	Good sources of magnesium are: fruits or vegetables, nuts, peas and beans, soy products, whole grains and milk.	-	are (P_1)	fruits or vegetables, nuts, peas and beans (O_1) soy products (O_1) whole grains and milk (O_1)	Good sources of magnesium (S_1)	-
2.	The RDAs for Mg are 300 mg for young women and 350 mg for young men.	-	are (P_1)	-	The RDAs for Mg (S_1)	300 mg for young women (O_1) 350 mg for young men (O_1)
3.	Increase potassium by ordering a salad, extra steamed or roasted vegetables, bean-based dishes fruit salads, and low-fat milk instead of soda.	-	-	salad (S_1) extra steamed or roasted vegetables (S_1) fruit salads (S_1) law-fat milk (S_2)	Increase potassium (S_1) soda (S_3)	-
4.	Babies need protein about 10 g a day.	Babies (S_1)	need (P_1)	-	protein (O_1)	10 g a day (O_1)
5.	1 teaspoon of table salt contains 2300 mg of sodium.	-	contains (P_1)	table salt (S_1)	sodium (O_1)	1 teaspoon (S_1) 2300 mg (O_1)
6.	Milk, cheese, yogurt and other dairy products are good sources of calcium and protein, plus many other vitamins and minerals.	-	are (P_1)	Milk, cheese, yogurt and other dietary products (S_1)	good sources of calcium and protein (O_1) many other vitamins and minerals (S_2)	-
7.	Breast milk provides sufficient zinc, 2 mg/day for the first 4-6 months of life.	-	provides (P_1)	Breast milk (S_1)	sufficient zinc (O_1)	2 mg/day for the first 4-6 months of life (O_1)
8.	If you're trying to get more omega-3, you might choose salmon, tuna, or eggs enriched with omega-3.	you (S_2)	're trying to get (P_2) enriched (P_3)	salmon, tuna, (O_2) eggs (S_3)	more omega-3 (O_2) omega-3 (O_3)	-
9.	If you need to get more fiber, look to beans, vegetables, nuts and legumes.	You (S_2)	need to get (P_2)	beans, vegetables, nuts, and legumes (O_2)	more fiber (O_2)	-
10.	Eating foods high in vitamin C and iron can reduce the absorption of ingested nickel.	-	can reduce (P_1)	Eating foods (S_1)	vitamin C and iron(S_1) the absorption of ingested nickel (O_1)	-
11.	The body of a 76 kg man contains about 12 kg of protein.	-	contains (P_1)	-	protein (O_1)	The body of a 76 kg man (S_1) about 12 kg (O_1)
12.	Excellent sources of alpha-linolenic acid, ALA, include flaxseeds and walnuts.	-	include (P_1)	flaxseeds and walnuts (O_1)	Excellent sources of alpha-linolenic acid (S_1) ALA(S_1)	-
13.	The recommended intake for total fiber for adults 50 years and younger is set at 38 g for men and 25 g for women, while for men and women over 50 it is 30 g and 21 g per day, respectively, due to decreased food consumption.	50 years (S_1) younger (S_1) for men and women over 50 (S_2) it (S_2)	is set (P_1) is (P_2)	decreased food consumption(S_3)	The recommended intake for total fiber for adults (S_1)	38 g for men (O_1) 25 g for women (O_1) 30 g (O_2) 21 g per day (O_2)
14.	I'm good at tennis.	-	-	-	-	-
15.	Your hat looks very nice.	-	-	-	-	-

If we look at sentence 5, "*1 teaspoon of table salt contains 2300 mg of sodium.*", the recommendation is not split by the splitting proposed in the pre-processing part because it does not consist of $ADVP$, $CONJP$, $SBAR$, or two successive chunks that are (O,O). By using the drNER method ("table salt", S_1) is the *Food* entity extracted. There is one *Action* entity, ("contains", P_1), one *Nutrient* entity, ("sodium", O_1), and two *Quantity/Unit* entities, ("1 teaspoon", S_1) and ("2300 mg", O_1). The labels that are given to each of the extracted entities are the labels for the *Subject*, *Predicate*, and *Object* that help us better to interpret the extracted information. For example, in this dietary recommendation we have two *Quantity/Unit* entities, one of them is related to an entity extracted in the subject of the sentence, and the other one is related to an entity that is extracted in the object of the sentence. For example, from the label of the *Quantity/Unit* entity, ("1 teaspoon", S_1), we can see that this entity is related to some other entity that is extracted from the same part of the sentence, the *Food* entity, ("table salt", S_1). The other *Quantity/Unit* entity, ("2300 mg", O_1), is related to an entity that is found in the object of the sentence, or in our case is related to the *Nutrient* entity, ("sodium", O_1). Finally, the extracted knowledge can be interpreted as (("1 teaspoon", S_1),("table salt", S_1)) and (("2300 mg", O_1),("sodium", O_1)), and ((("1 teaspoon", S_1),("table salt", S_1)); ("contains", P_1); (("2300 mg", O_1),("sodium", O_1))).

Alternatively, the sentence 13, "*The recommended intake for total fiber for adults 50 years and younger is set at 38 g for men and 25 g for women, while for men and women over 50 it is 30 g and 21 g per day, respectively, due to decreased food consumption.*", consists of two adverb chunks, so it needs to be split. If the recommendation is not split, then the entities "The recommended intake for adults" and "50 years" are extracted as *Group* entities, "is set" and "is" are extracted as *Action* entities, "decreased food consumption" is extracted as *Food* entity, and "38 g for men" and "25 g for women" are extracted as *Quantity/Unit* entities, and the information for men and women over 50 remains hidden, since it is not extracted. For this reason, the recommendation is split in the location of each adverb chunk. In this recommendation, there are two adverb chunks, "while" and "respectively", so we split it in three parts, "*The recommended intake for total fiber for adults 50 years and younger is set at 38 g for men and 25 g for women*", "*For men and women over 50 it is 30 g and 21 g per day*", and "*Due to decreased food consumption*". The proposed method is then used on each part of the recommendation obtained after splitting. For the first part, the extracted entities are: ("The recommended intake for total fiber for adults", S_1) as *Nutrient* entity, ("50 years", S_1) and ("younger", S_1) as *Group* entities, ("is set", P_1) as *Action* entity, and ("38 g for men", O_1) and ("25 g for women", O_1) as *Quantity/Unit* entities. By applying the method on the second part of the recommendation, the extracted terms are ("For men and women over 50", S_2) and ("it", S_2) as *Group* entities, ("is", P_2) as *Action* entity, and ("30 g", O_2) and ("21 g per day", O_2) as *Quantity/Unit* entities. For the third part of the recommendation, only one extracted term exists ("decreased food consumption", S_3).

4.4.3 Evaluation

Test corpus

Due to the lack of annotated corpora in the domain of dietary information and in order to evaluate the newly proposed NER method for evidence-based dietary recommendations we created a test corpus. The main question was how to select the documents for the test corpus. We fixed the number at 100 because after extraction we need to manually check the extracted information. In order to promote diversity in the test corpus, we selected the documents from heterogeneous sources. We did this because different heterogeneous

sources have different ways of reporting dietary recommendations. Fifty documents are dietary recommendation summaries, which are extracted from the scientifically validated web site, the Food and Nutrition Information Center of United States Department of Agriculture [139]. These documents are extracted from 12 different institutions and their distribution per institution is presented in Figure 4.5.

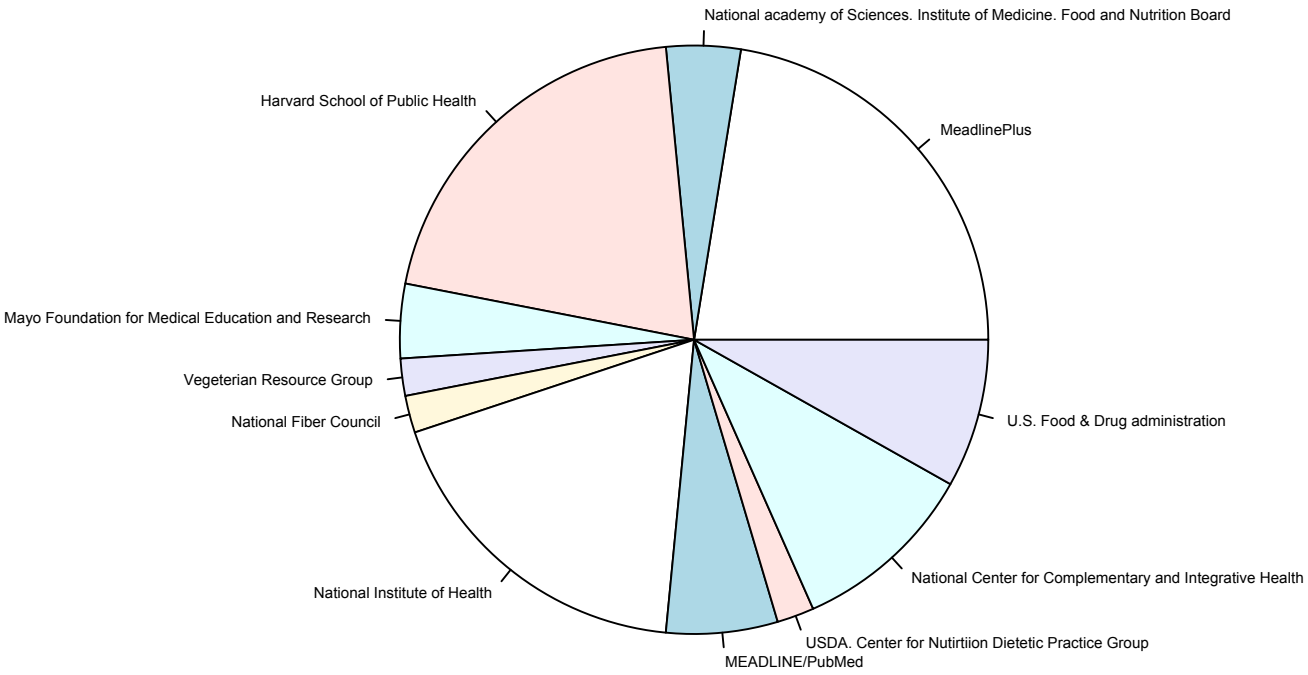


Fig 4.5: Distribution of documents per institution.

The other 50 documents are abstracts of scientific publications. They were selected by using the PubMed API [183] in combination with two keywords, "food composition" and "dietary intake". Twenty five abstracts were selected randomly from the documents that were returned for each key word. The average number of sentences per document is 3.88 for documents from scientifically validated web sites and 6.54 for abstracts of scientific publications. This result is reasonable because the paragraphs from the scientifically validated web sites are summaries of dietary recommendations, while the abstracts from the scientific publications may contain dietary recommendations, but also contain other information about the study. The distribution of the number of sentences per document for the two subsets of the test corpora is presented in Figures 4.6 and 4.7.

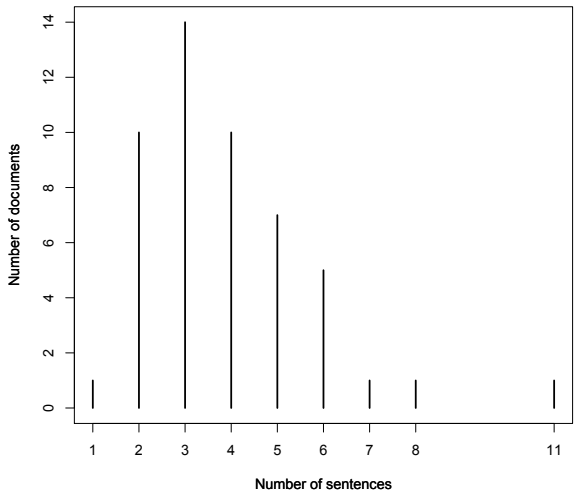


Fig 4.6: Distribution of number of sentences per documents from scientifically validated web sites.

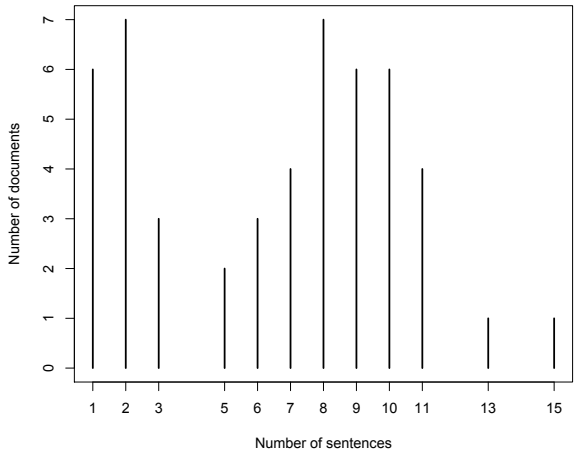


Fig 4.7: Distribution of number of sentences per documents that are abstracts from scientific publications.

This test corpus is not annotated. The results that are obtained by the drNER are further manually checked by two human experts, who are clinical dietitians, in order to see if the extracted entities have the correct label and if there are missing entities by drNER. The test corpus together with the obtained results for each document, separately, are available at the following link dx.doi.org/10.17504/protocols.io.hqbb5sn.

Results

We evaluated the drNER on the above test corpus. For each document in the test corpus, we tried to extract all useful information related to dietary recommendations, with a focus on the *Food* entities, *Nutrient* entities, and *Quantity/Unit* entities. In Table 4.9, the

result of the evaluation for each entity is presented. In this table, the results for the documents that are extracted from scientifically validated web sites and scientific publications are provided, separately. At the bottom of the table the summary of results of our test corpus is presented. The results are presented by reporting the number of true positives, false positives, and false negatives. The true positives are the extracted entities for which the obtained entity label from drNER and the human expert is the same. The false positives are the extracted entities for which the drNER label (known as expectation) is positive, but this is false according to the external judgment of the human expert. For example, for the *Food* label, if an extracted entity is labeled as *Food* by drNER, but the human expert provided that the true label is *Nutrient*, it means that this is false positive for the *Food* label. The false negatives are the entities for which the drNER label (known as expectation) is negative, but this is false according to the external judgment of the human expert. For example, for the *Food* label, if an extracted entity is not labeled as *Food* by drNER, but the human expert provided that the true label is *Food*, it means that this is false negative for the *Food* label. Also, in false negatives are some entities that are not recognized as a given entity of interest by drNER, but the human experts assumed that this information from the document should be extracted.

Table 4.9: Evaluation results.

	Food			**Nutrient**			**Quantity/Unit**		
	TP	**FP**	**FN**	**TP**	**FP**	**FN**	**TP**	**FP**	**FN**
Scientifically validated web sites	326	5	22	243	0	13	47	0	2
Scientific publications	213	0	3	314	2	4	39	0	9
Total	539	5	25	557	2	17	86	0	11

TP indicates true positives
FP indicates false positives
FN indicates false negatives

The number of true positive *Food* entities is 539. Out of them 326 are from the documents that are extracted from scientifically validated web sites and 213 are from scientific publications. Also, there are 5 false positives, from the scientifically validated web sites, related to phrases that consist of verbs related to food such as "need to eat", "eating", etc., when they are not selected as an *Action* entity. In the future, this can be omitted by checking the tags of the chunks, so if they are verb chunks they can be removed from the candidate solutions. The number of false negatives for the *Food* entity is 25, out of which 22 are from the documents from the scientifically validated web sites and 3 of them are found in the scientific publications. They occur because "grains" is not recognized as a *Food* item by the USAS English semantic tagger we used as our dictionary. For some false negatives when the information about the *Food* entity is not extracted because it is not found in the dictionary, the information about it can be extracted in the *Group* entity that is proposed by the method to catch some additional useful information related to dietary information.

In the case of the *Nutrient* entity, the number of true positives in our test corpora is 557. Out of them, 243 are found from the documents extracted from the scientifically validated web sites and 314 are extracted from scientific publications. The number of false positives is 2, and the number of false negatives is 17. Out of them, 13 are extracted

from the documents from scientifically validated web sites and 4 are found in the scientific publications. Most of them are related to the fact that none of the corpus-based NER, that we used as dictionaries, recognize the concepts as a chemical entity. Some of them are related to "omega-3s". It is interesting that "omega-3" is recognized, but the plural form is problematic for all dictionaries we used. In the future, these results can be improved by adapting a heuristic approach for linking the tokens to the dictionaries.

For the *Quantity/Unit* entity, the number of true positives is 86. Out of them, 47 are from the documents extracted from the scientifically validated web sites and 39 of them are found in the scientific publications. We did not find any false positives and the number of false negatives is 11. The false negatives are related to some units such as "ngg(-1)", "mgkg", etc. This happens because these units do not exist in the dictionaries we used. Also, we use lemma of each token when we link it to the dictionaries related to the *Quantity/Unit* entity in order to distinguish between singular and plural forms of the units. We did this only for the *Quantity/Unit* entity because for others we used corpus-based NER systems that already include this information.

In Table 4.10 results of the performance measures of drNER for each entity obtained on our test corpus are presented. The good result obtained for the precision was expected because the false positives are very rare, since the chunk will be a candidate for some entity only if some of its tokens is found in the entity dictionary. From the results obtained for the recall, we can conclude that the recall for the *Quantity/Unit* entity is the lowest. This was expected because in this case the dictionary we used was a combination of a standardized ontology of measurement units with a set of measurement units used for recipes. This linking can be assumed as pure terminological because each chunk token is linked to a dictionary that contains words that are measurement units. So if the token appears in the dictionary it will be returned, while if not, it will be missed and returned as false negative. From the other side, the proportion of false negatives in the case of *Food* entity and *Nutrient* entity is smaller compared to the proportion found for the *Quantity/Unit* entity, which was expected because for both we used a combination of corpus-based NERs as dictionaries that are obtained by using ML approaches and include orthographical, morphological, NLP, domain knowledge, and local context features. At the bottom of the table, the result for F_1 score for each entity is presented, which is a measure of accuracy or a weighted average of the precision and recall.

Table 4.10: Performance measures.

Performance measures	**Food**	**Nutrient**	**Quantity/Unit**
Precision	0.99	1.00	1.00
Recall	0.95	0.97	0.91
F_1 score	0.97	0.97	0.95

4.5 Discussion and Conclusion on drNER

To extract information from text, different NER methods can be used. Most of them are corpus-based NERs, while the others are rule-based, where rules are created using the characteristics of the entities of interest. The first part of our second hypothesis is that, to extract information from an untapped domain, a rule-based NER method can be used without an annotated corpus, where the rules will not be associated to the characteristics

of the entities of interests. The hypothesis is confirmed in a dietary domain, especially information extraction from evidence-based dietary recommendations. Experimental results have shown that promising results can be achieved.

To the best of our knowledge, drNER is the first NER method that is focused on information extraction of evidence-based dietary recommendations [30]–[32]. The dietary domain brings a new application domain, with similar goals as previous IE shared tasks on biological event extraction. However, an annotated corpus does not exist, so there are no methods that focus on information extraction. Because of that, a comparison of the drNER is made with some NER methods that can be used for each entity, separately, or they are NER methods that can in our case be used as dictionaries for some entities. For example, let us focus on one sentence "*People of any age who are African Americans should further reduce sodium intake to 300 mg per day.*" [139]. By using the USAS English online tagger and becas[chemicals] API, the *Nutrient* entity which will be extracted is "sodium", while by using the becas API it is not recognized. By using the dictionaries applied for the *Quantity/Unit* entity, the only entity extracted here will be "mg". The result by applying the drNER on this sentence is ("People of any age", S_1) and ("who are African Americans", S_1) as *Group* entities, ("should further reduce", P_1) as *Action* entity, ("sodium intake", O_1) as *Nutrient* entity, and ("300 mg per day", O_1) as *Quantity/Unit* entity. If the recommendation is *"The RDAs for Mg are 300 mg for young women and 350 mg for young men."*, by using the USAS English online tagger the result for the *Nutrient* entity is *"Mg"*, while by using the drNER it is *"The RDAs for Mg"*. So instead of extracting only the nutrient component, the drNER could also extract the type of the DRVs reported. In our proposed method, by applying the three proposed post-hoc chunkings, we can obtain also the phrases that differ from the phrases that can be obtained by the corpus-based NERs used as dictionaries and give us more information for the entities. Also, adding the *Action* entity and the labels for the *Subject*, *Predicate*, and *Object* provides additional information, and the *Group* entity helps catch information that could be important to better interpret the extracted information.

To compare the methodology used in the drNER, we compare it with methodologies used by other NER methods used for other biomedical domains. For example, the SeeDev task that was a part of the 4th BioNLP Shared task consists two subtasks, SeeDev-binary on binary relation extraction and SeeDev-full on full event extraction. Because there are annotated corpora, all seven teams used supervised ML approaches. Five systems used SVMs and two systems were based on different algorithms, maximum entropy (MaxEnt) and a convolutional neural network. The methodology of the drNER method is completely different from the methodologies used by these approaches. The drNER is a pure NLP method that is not based on annotated corpora, while all the SeeDev methods are based on ML approaches. Also, the entities involved in drNER are related to the dietary domain, and the SeeDev approaches involved in the extraction are related to genetic and molecular mechanisms involved in plant seed development. Another method is BANNER, evaluated on the BioCreative 2 GM training corpora, which is designed to maximize domain independence by not employing semantic features or rule-based processing steps. The domain-specific performance is not the purpose of the system, but researchers could adopt BANNER for a specific domain, by applying two types of post-processing. BANNER is based on annotated corpora and it uses CRFs. So the difference between the methodology used by BANNER and drNER is the same as the difference in the approaches presented on SeeDev task. Because an annotated corpus in the dietary domain does not exist, it is better to compare the methodology of the drNER with the methodologies of rule-based NERs that exist in the biomedical domains. Many rule-based NER methods use rules that combine terminological resources and the characteristics of the entities, but to write rules

that depend from the characteristics for each entity is a time-consuming task. Further more, it requires a good understanding of the domain. For example, Lowe et al. [184] give a grammar and dictionary driven approach to entity recognition that uses a mixture of expertly curated grammars and dictionaries , as well as dictionaries automatically derived from public resources. They have created 486 rules. The benefit of this approach is that it works well when you do not have annotated data but requires dictionaries and grammars related to each entity.

The benefit of using drNER is that it can provide promising results even when an annotated corpus is missing. The construction of the annotated corpus for a new domain is a time-consuming task and requires effort by human experts to produce it. Also, rule-based NERs require expertly created grammars and dictionaries that are further used for information extraction. Creating such grammars for a new domain is also a time-consuming task and requires human experts to produce them. In our case, we do not have expertly created grammars for the entities we are interested in, but we only used a small number of Boolean algebra rules that are unrelated to the characteristics of the entities, but help us to automatically define the phrases that are entities mentions. The time complexity of the current version of the code, which is non-optimized and is developed only for testing propose of the methodology is in average of 10s per sentence. Further, using parallel programming and optimization of the code, this can be improved.

The main disadvantage of the method is that it uses dictionaries for each entity, so only the entity mentions that exist in the dictionaries will be recognized. However, the benefit of using such a method is related to the frequent updates of such dictionaries including new entities with their synonyms. Also, in our case, we do not use only classical terminological dictionaries, but we allow using corpus-based NERs for some entities of interest as dictionaries, if they exist previously from the literature.

drNER is a software developed in the R programming language. The code is publicly available on GitHub (https://github.com/teftimov/drNER-Git.git) where the two phases and the steps involved in each one are programmed as separate functions.

We present a NER method for knowledge extraction of evidence-based dietary recommendations, called drNER. The goal of this method is to promote progress in information extraction in the field of dietary domain, especially focused on three main entities: *Food*, *Nutrient*, and *Quantity/Unit*. The dietary domain brings a new application domain, which has similar tasks on biomedical extraction, and is crucial for promoting health and well-being.

The proposed NER method for information extraction of evidence-based dietary recommendations is a combination of terminological-driven NER and rule-based NER. The difference with the purely terminological-driven NER methods is that we allow for the use of corpus-based NERs as dictionaries for some entities of interest, instead of using dictionaries that consist of concepts and synonyms. The difference with the rule-based NERs is that we do not use rules based on the characteristics of the entities. We only have a small number of Boolean algebra rules that are not related to the characteristics, but help us to define the phrases that are entity mentions. The method consists of two-phases. The first phase involves the detection and determination of the entity mentions. It works by using some NLP methods and linking each token to a dictionary for each entity in which we are interested. After that, it uses three post-hoc chunkings in order to better determine the entities mentions. The second phase is the selection and extraction of the entities. It is based on text syntactic analysis. Finally, by applying the rules defined in this phase, we can extract useful information related to dietary recommendations.

To the best of our knowledge, drNER is the first NER method where the focus is in the domain of evidence-based dietary recommendations, which is an untapped domain.

The evaluation of drNER is done on test corpus that includes 100 documents. The best results achieved rely on the fact that for some entities such as *Food* and *Nutrient*, the terminological resources are not classic dictionaries that consist of concepts with synonyms, but they could be some corpus-based NERs that exist. For example, the *Nutrient* entity is related to chemical-named entity recognition. By using chemical NERs we can obtain the chemical information, but the type of DRVs or some additional information associated with it, is not extracted. For this purpose, three post-hoc chunkings are presented and help in modelling the dietary domain.

For future work, we plan to normalize the extracted entities. Then, we will try to find a good way to represent the extracted knowledge to human experts. By using the extracted knowledge from the dietary domain and the knowledge for drugs, diseases, and genes, that can be obtained from methods presented as a part of shared workshops, we will try to build an annotated corpus and to increase adoption of linked data techniques as an effective solution to knowledge representation and management in Life and Health Sciences [185]. Having an annotated corpus and knowledge representation, the next step will be to extract the relations that exists between these entities by applying AL or corpus-based NERs.

Chapter 5

String Similarity of Domain-Specific Short Text Segments

An extracted concept, which is a result of information extraction, may be mentioned using phrases with a variety of structures, which is a consequence of how people express themselves. Sometimes, this information for the same concept that is represented on different ways needs to be combined for further analysis and one way to do this is to apply text normalization methods. Text normalization methods are based on string similarity measures that can fail when they are dealing with short segments of text. In this chapter, we explain the problem of string similarity of domain-specific short segments of text, working on a food matching problem. We provide an overview of standard string similarity measures. We present a methodology that uses a probability theory to model the domain using the morphological information that is presented in the text. Finally, we present the experimental results together with the discussion and the benefits of using the proposed methodology.

5.1 Problem Definition

The extracted information from text, which is a subject of our interest, can be mentioned using phrases with a variety of structures. The different representations of the same concept creates a problem to the automatic integration of information. For this reason, normalization methods are used, in which the main purpose is to link a concept mention to a known concept from a domain-specific terminological resource. Normalization methods are based on string similarity measures.

Measuring string similarity plays an increasingly important role in text related research. String similarity measures operate on string sequences and give us a metric of similarity (or dissimilarity) between two text strings. They can be performed on a character or a term level. Working with extracted information from a given domain, the concepts are in the most cases described as short segments of text. Metzler et al. [108] presented a study of measuring the similarity between short segments of text. They analyzed various types of text representations (surface, stemmed, and expanded) in a combination with several similarity measures, including lexical matching and probabilistic measures based on language models. They have shown that lexical matching is good for finding semantically identical matching, while probabilistic models are better at finding topically related matches. They also presented that standard string similarity measures can fail when tasked with computing the similarity between two very short segments of text.

To help text normalization methods, especially when they are applied on short segments of text in a given domain, we propose a novel method, which introduces a probability model

of the domain using the morphological information presented in the text. We consider a real problem on food matching, which is actually a subject undergoing intense study in food science. In 2016, EuroFIR organized a workshop on food matching at the EuroFIR Food Forum 2016 and in 2017, food matching was a special topic covered at the 12^{th} International Food Data Conference. To help the automatic transformation from food intake to nutrient intake, the problem that appears is that the name of the same food (food concept) can be different regarding different ways of how people express themselves. For example, we can have "chicken breast, raw" that needs to be matched to a FCDB in order to find its nutrition profile, and the FCDB consists of "chicken breast, cooked, salted", "raw chicken breast", and "chicken breast". On the performed matching depends how accurate will be the calculated nutritional value. One way to perform food matching is to map these descriptions to a concept that exists in a terminological resource, a food classification and description system. By mapping them, they obtain unique identifiers, which help the process of combining the information for some concept.

5.2 Related Work

Normalization methods are based on string matching functions, where the idea is to find the best possible matches. For this reason, string similarity measures are typically calculated and can further be combined using some heuristics in order to find the most relevant match for a given concept.

In Chapter 2, we explained that string similarity measures can be applied on the character and term level [107], [108]. These metrics also depend on the textual representation and they may fail when they are used for string similarity of short segments of text [108].

Normalization methods that are based on string similarity measures are presented in [85], [186], [187]. Also, there are methods that use post-processing rules applied after the concept is extracted or regular expressions to find some matches for which a specific form may not occur in the terminological resource [188]. Some normalization methods are based on ranking technique in order to rank the candidate matches and then to find the most relevant one [189]. An example of automatic normalization method focussed on phenotypic information, which integrates a number of different similarity measures, is presented in [190]. Normalization methods can also use ML algorithms to improve results, which was shown in the gene normalization task as part of BioCreative II [191] and BioCreative III [192]. Tsuruoka et al. [193] used logistic regression for learning a string similarity measures for gene/protein name dictionary look-up. Leaman et al. [194] introduced DNorm, which is a technique for disease normalization based on pairwise learning to rank

Another method of publishing structured data that can be interlinked and become more useful through semantic queries is linked data. It is based on standard Web technologies and it extends them to share information in a way that can be read automatically by computers. It follows the idea of Semantic Web, where the core element is an ontology. An ontology is a formal explicit specification of a shared conceptualization [195]. Boulos et al.[196] provided a review of the existing food ontologies that can be used with some modifications and together with some relevant non-food ontologies. This is in order to enable users to make the correct, healthy food-and-drink choices that are personalized for their particular health condition, age, body weight, lifestyle and preferences, and also give their coverage gaps, the incompleteness or limited scope, which are challenges for developing an universal, comprehensive food ontology. The presented food ontologies are: FoodWiki [197], AGROVOC [198], Open Food Facts [199], Food Product Ontology [200], and FOODS (Diabetics Edition) [201]. However, an ontology for nutrition science that will be more comprehensive is a task that is currently going under the H2020 Richfields

project. For this reason, we are not focusing on linked data methods, but only on text normalization methods based on string similarity measures.

To the best of our knowledge, no previously reported automatic normalization method has focussed on the food and nutrition domain. In 2016 and 2017, food science puts a special focus on food matching, or mapping food consumption to food composition data. In 2016, EuroFIR, organized a workshop on food matching at the EuroFIR Food Forum 2016, where different European countries presented their results and experience of matching food items. All of the proposed methods were based on manually matching the food items to a food information terminological resource, which is a time-consuming task. In 2017, food matching was a special topic covered at the 12^{th} International Food Data Conference, where different groups presented semi-automatic approaches for food matching based on standard string similarity measures, but there are no published papers yet.

5.3 Methodology

To improve string similarity measures of domain-specific short text segments, a novel method for string similarity of domain-specific short text segments called POS tagging probability weighted method is proposed [33]–[35], [37], [38]. The name comes because the method uses the morphological information provided by the POS tagging in combination with the probability theory. It consists of two steps. The first step is to use POS tagging to obtain the morphological information presented in the text. The second step is to define which of the extracted morphological information is relevant for the domain of interest and then to use the probability theory to model the domain using the relevant morphological information.

Further, we continue by explaining the idea behind the proposed method in general. Then we give an explanation of its application in the food domain, especially the food matching problem.

5.3.1 Part-of-speech tagging probability weighted method

Let D_1 and D_2 be two short segments of text. First POS tagging, also called grammatical tagging, is applied on each of them to identify the part-of-speech tags such as nouns (NN, NNS, NNP, NNPS), verbs (VB, VBD, VBG, VBN, VBP, VBZ), adjectives (JJ, JJR, JJS), cardinal numbers (CD), etc. Let us define

$$Y_i = \{tokens\ from\ D_i\ that\ belong\ to\ one\ word\ class\}, \tag{5.1}$$

where $i = 1, 2$. The word classes are: nouns, adjectives, verbs, adverbs, prepositions, determiners, pronouns, conjunctions, modal verbs, particles, and numerals. For example, Y_i can be a set of all tokens from D_i that are tagged as nouns. In such case, the set consists of all tokens that are tagged as NN, NNS, NNP, and NNPS.

The next step is to define which of the extracted word classes (morphological POS tags) are significant to describe the domain in which the short segments of text belong. The set of nouns is crucial because nouns carry most of the information in the text, while all other word classes (adjectives, verbs, numbers, etc.) only give an additional explanation. After extracting the set of nouns and the sets of other word classes that are significant for the domain, lemmatization is applied on each of them. To find string similarity between both short segments of text, a probability event is defined as a product of independent events

$$X = N \prod_{j=1}^{k} Z_j, \tag{5.2}$$

where N is the similarity between the sets of nouns found in both short segments of text, k is the number of additional word classes that are selected and are significant for the domain, and Z_j is the similarity between the sets of word class, j, found in both short segments of text. The additional word classes can be adjectives, verbs, etc.

Because these events are independent, the probability of the event X can be calculated as

$$P(X) = P(N) \prod_{j=1}^{k} P(Z_j). \tag{5.3}$$

To calculate it, the probabilities of the independent events need to be defined. Because the problem looks for the similarity between two sets, it is logical to use the Jaccard index, J, which is used in statistics for comparing similarity and diversity of sample sets [202]. For the similarity between the nouns, the Jaccard index is used, while for the similarity between the additional word classes the Jaccard index in combination with Laplace probability estimate [203]; this is because, in some short segments of text, the additional information provided by other word classes can be missed, so there will be no zero probabilities. The probabilities are calculated as

$$\begin{aligned} P(N) &= \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}, \\ P(Z_j) &= \frac{|Z_{j_1} \cap Z_{j_2}| + 1}{|Z_{j_1} \cup Z_{j_2}| + 2}. \end{aligned} \tag{5.4}$$

By substituting Equations 5.4 into Equation 5.3, we obtain a weight for the matching pair.

If we need to map a concept described by a short segment of text to a concept that exists in a terminological resource, first POS tagging is used on its description to define the sets of nouns and the additional word classes that are significant for the domain. On each extracted set, lemmatization is applied. The short segments of text that describe the concepts that exist in the terminological resource are then searched for all extracted nouns and each one that consists of at least one of the provided nouns is returned. The result is a subset of concepts that exist in the terminological resource and their short segments of text consist of at least one of the nouns obtained from the concept that needs to be mapped. Then, for each concept in the subset, POS tagging is used to extract the nouns and the additional word classes, which are further processed by lemmatization. The next step is to define the similarities between the concept that needs to be mapped and the concepts that belong to the subset. To do this, the probability of the similarity event is calculated between the concept that needs to be mapped and each of the concepts that belongs to the subset. The matching pair with the highest probability is the most relevant found match.

If we focus on food domain, or especially the food matching problem, let D_1 and and D_2 be names of two food items. As we said before, the nouns carry of most of the information, while the additional word classes that describe the food domain are adjectives, which explain the food item in the most specific form (e.g., frozen, fresh), and the verbs, which are generally related with the method of preparation (e.g., cooked, drained). Let us define

$$\begin{aligned} N_i &= \{nouns\ extracted\ from\ D_i\}, \\ A_i &= \{adjectives\ extracted\ from\ D_i\}, \\ V_i &= \{verbs\ extracted\ from\ D_i\}, \end{aligned} \tag{5.5}$$

where $i = 1, 2$.

To find the similarity between these two names of food items, an event is defined as a product of two other events

$$X = N \cdot (A + V), \tag{5.6}$$

where N is the similarity between the nouns found in N_1 and N_2, and $A+V$ is the similarity between the two sets of adjectives and verbs handled together as $A_1 + V_1$ and $A_2 + V_2$. The adjectives and verbs are handled together to avoid different forms with the same meaning. For example, if adjectives and verbs are handled separately, the match "apple dry" and "dried apples" will not be a perfect match. However, for the computer "dry" and "dried" are completely different words that have the same meaning. The same is also true for the singular and the plural form of "apple". Also, if adjectives and verbs are handled separately, the match "browned bread" and "brown bread" will not be a perfect match because "browned" is a verb and "brown" is an adjective, but both have the same meaning. To avoid this, lemmatization is applied for each extracted noun, verb and adjective, and the similarity event uses their lemmas.

Because these two events are independent, the probability of the event X can be calculated as

$$P(X) = P(N) \cdot P(A + V). \tag{5.7}$$

The probabilities are calculated as

$$\begin{aligned} P(N) &= \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}, \\ P(A + V) &= \frac{|(A_1 \cup V_1) \cap (A_2 \cup N_2)| + 1}{|(A_1 \cup V_1) \cup (A_2 \cup N_2)| + 2}. \end{aligned} \tag{5.8}$$

By substituting Equations 5.8 into Equation 5.7, we obtain a weight for each matching pair.

5.4 Experimental Results

To evaluate the proposed method, we performed two experiments in the food domain. In the first one, we compare the food matching results obtained using the proposed method and some standard string similarity measures, while in the second one we used it as a part of system that can be used to standardize foods according to a food classification and description system. For both experiments we tried to map foods from Slovenian national databases to food concepts that exist in a terminological resource, which in our case is a food classification and description system called FoodEx2.

5.4.1 FoodEx2

In 2011, the European Food Safety Authority, EFSA [3], introduced a comprehensive food classification and description system for exposure assessment, known as FoodEx1 [204], aimed at covering the need to describe food in data collections across different food safety domains. After a testing phase, in 2015, EFSA introduced a new version called FoodEx2 [204], in order to match the needs expressed by different users. The system consists of a lot of individual food items aggregated into food groups and broader food categories organized in a hierarchical relationship. In addition, it provides generic food descriptions that represent the minimum level of detail required for making intake or exposure assessments. The description is provided using facets, which are a collection of terms that describe the properties and aspects of foods from various perspectives.

FoodEx2 is applicable across different domains, in particular food consumption, chemical contaminants, pesticide residues, zoonoses, and food composition. It allows new forms

of dietary assessment technology, and connects food intake and food composition data enabling the conversion of food intake into nutrient intake [205]. The system consists of three types of food categories representing three different levels in the food chain, which involve an increasing level of food processing i.e. moving from raw commodities to derivatives to composite foods. Raw commodities (r) are the parts physically separated from a living source after harvesting (plants) or slaughtering (animals) and the processes it assumes to be involved, in the context of this classification, do not change the nature of food. Derivatives/ingredients (d) are food products obtained from raw commodities by applying a process that changes the nature of food. Composite foods are obtained by using more raw commodities and/or derivatives through processes that always involve recipes, simple composite food (s) and aggregated composite food (c). For example, simple composite foods include muesli, porridge, cereal bars, and jam, while examples of aggregated composite foods include pasta-based dishes, cakes, croissants, soups and salads. This means that processing has a critical role in defining whether a specific food item is raw, a derivative or a composite. Such processes are defined in FoodEx2 technical report [204].

FoodEx2 includes two types of terms: list terms and facet descriptors. List terms are terms defining a food group, or, in a broader sense, taking into account the expanded scope of FoodEx2, a defined matrix for data reported in the food safety domain. List terms are the food groups appearing in the tree structure of the different hierarchies and may have a different level of aggregation. Depending on the scope of the groups, they may be reportable or not; in other words, some list terms are suitable for describing a food item, while coding data and other list terms are only suitable for browsing a hierarchy or creating a summary table for data analysis. The list terms that can be used to describe a food item are the base terms, because these can form the basis of a more detailed and complex FoodEx2 code for data reporting. List terms are represented by a code, for example, the A0C75 is the code for "salmon". Facet descriptors are elements of additional information included in or added to the list terms, each providing different options to describe a particular aspect of a food category, such as treatments received, production method, fat content, and qualitative information. FoodEx2 consists of 32 facets that can be either implicit or added. Implicit facets are applied to a specific food category, while added facets describe a characteristic of a specific food item. Each food category (r, d, s, or c) may have different facet descriptors. FoodEx2 code includes different pieces of coded information in a single string, which contains a code for a list term (mandatory), followed by a hashtag "#" and an unlimited sequence of facets, each separated by dollar characters "$". For example, if the food item "nectar, orange" needs to be coded with FoodEx2, the code already existed in FoodEx2 and is A03BG. If additional information in the food description is presented, for example that this nectar is sugar free, organic, and fortified with calcium, the FoodEx2 code looks like this: A03BG#F09.A0EXH$F10.A077L$F21.A07SE. This code contains the information that this is nectar, orange (A03BG), FORTIFICATION AGENT = Calcium (F09.A0EXH), QUALITATIVE INFO = Sugar free (F10.A077L), PRODUCTION METHOD = Organic production (F21.A07SE) [204]. Using FoodEx2 is not difficult, but the coding is a process that is always based on a choice by a person and needs to be manually performed.

5.4.2 Comparing the proposed method with some standard string similarity measures

To evaluate the proposed method, first we compared the results obtained using it with the results obtained with some standard string similarity measures. For this reason, we used the English names of 532 food items from the Slovenian national FCDB and for each one we tried to find the most relevant match that exists in FoodEx2. From the available FoodEx2

data [204], 5,416 instances of food items were selected. Instances are food items that have the attribute "Statef" with the value r, d, s, or c. This attribute indicates the level of the food category represented by the term in the food chain e.g., a raw, a derivative, a simple composite, or an aggregated composite food.

Let D_1 and D_2 be two short segments of text. We selected nine measures as standard string similarity measures:

- The *Levenshtein distance* counts the number of deletions, insertions and substitutions necessary to turn D_1 into D_2.
- The *Optimal String Alignment distance* is like the Levenshtein distance but also allows transposition of adjacent characters. Each substring may be edited only once.
- The full *Damerau-Levenshtein distance* is like the optimal string alignment distance except that it allows for multiple edits on substrings.
- The *longest common substring* is defined as the longest string that can be obtained by pairing characters from D_1 and D_2 while keeping the order of characters intact.
- A *q-gram* is a subsequence of q consecutive characters of a string. If x (y) is the vector of counts of q-gram occurrences in D_1 (D_2), the q-gram distance is given by the sum over the absolute differences $|x_i - y_i|$.
- The *cosine distance* is computed as $1 - \frac{x \cdot y}{||x||||y||}$, where x and y were defined above.
- Let X be the set of unique q-grams in D_1 and Y the set of unique q-grams in D_2. The *Jaccard distance* is defined as $1 - \frac{|X \cap Y|}{|X \cup Y|}$.
- The *Jaro distance* is defined as $1 - \frac{1}{3}(w_1 \frac{m}{|D_1|} + w_2 \frac{m}{|D_2|} + w_3 \frac{(m-t)}{m})$, where $|D_i|$ indicates the number of characters in D_i, m is the number of character matches and t the number of transpositions of matching characters. The w_i are weights associated with the characters in D_1, characters in D_2 and with transpositions.
- The *Jaro-Winkler distance* is a correction of the Jaro distance. It uses a prefix scale p which gives more favourable ratings to strings that match from the beginning for a set prefix length l.
- The *skip-grams* are generalization of *n-grams* in which the components (typically words) need not be consecutive in the text, but may leave gaps that are skipped over.

These string similarity measures are performed using the R package "stringdist". More details about how to use this package are available in the "stringdist" package for approximate string matching [206]. The skip-grams are performed using the R package "tokenizers" [207].

To compare different string similarity measures, we manually labeled all the instances that are results of food matching. The labels assigned are:

- 0 - either no match is found or the match found is not the correct one;
- 1 - multiple matches are found with none of them being suitable or correct;
- 2 - multiple matches are found in which one of them is the correct one;
- 3 - perfect match is found.

Table 5.1: Food matching results.

String similarity measure	0	1	2	3
Levenshtein	145	74	92	221
Optimal String Alignment	155	75	91	211
Damerau-Levenshtein	155	81	83	213
Longest Common Substring	152	54	67	259
q-gram, q=2	115	37	65	315
Cosine	289	9	13	221
Jaccard	242	84	34	172
Jaro	243	3	22	264
Jaro-Winkler	244	4	13	271
Skip-grams (bigrams with maximum skip distance of one)	88	161	68	214
POS tagging probability	58	4	**128**	**341**

Table 5.1 gives the food matching results for nine string similarity measures. The results are classified in the 4 above described labels.

From Table 5.1, it follows that POS tagging probability weighted method gives the most promising results. It gives 341 perfect matches, 128 matches that consist the correct match, or overall 88.35% correct food matches using the food items that belong to the 2 and 3 label. The matches that belong to the 2 label can be further processed with Jaro-Winkler, q-gram or other string similarity measure in order to obtain the correct match.

5.4.3 Standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2

There are two important rules when using FoodEx2. The first is to have prior knowledge of the system and the second is to choose the right starting point or what type of food it is with regard to FoodEx2 classification, i.e. is it a raw, a derivative or a composite food. This is important because different food categories may have different facet descriptors. By using FoodEx2 to classify and describe foods, it makes it easier to compare data from different data sources and to perform more detailed data analysis [208], [209]. Unfortunately, many food composition data and food consumption data, which need to be connected, are lacking since the process of classifying and describing food is performed manually, which is time-consuming and requires a good knowledge of the system. In 2016, EuroFIR [6] organized a workshop for food matching at the EuroFIR Food Forum 2016, where different European countries presented their results and experience of matching food items. Most countries used FoodEx2 codes, and all of the proposed methods were based on manually searching the FoodEx2 data, which is a time-consuming task. Among those that did not use FoodEx2 were a group from ETH Zurich, which presented a semi-automatic way of matching food items using food names and Jaccard distance applied at the character level, and our group, which presented a promising method for matching food items to food composition data, despite not being developed to match food items using FoodEx2.

In this experiment, we introduce a semi-automatic system, called StandFood, to standardize foods according to FoodEx2. The system consists of three parts. The first identifies what type of food is being analyzed (r, d, s, or c). This is the classification part that involves a machine learning (ML) approach. The second describes the food using the POS tagging probability weighted method, which results in the list term or FoodEx2 code for the food. The third combines the result from the first and the second part by defining

post-processing rules in order to improve the result for the classification part.

FoodEx2 data

From the available FoodEx2 data [204], the same 5,416 food items as in the first experiment were selected. Instances are food items that have the attribute "Statef" with the value r, d, s, or c. These were selected because before describing a food item, we need to determine to which food category a food item belongs.

StandFood

StandFood is a semi-automatic system for classifying and describing foods according to FoodEx2 [38]. It consists of three parts. The first one classifies foods into four FoodEx2 categories (groups), two for single foods: raw (r), derivatives (d), and two for composite foods, simple (s) and aggregated (c). For this purpose, it uses a ML approach. The second part is used for describing foods using the FoodEx2 facets, by using a POS tagging probability weighted method. The third part combines the result from the first and the second part by defining post-processing rules in order to improve the result for the classification part.

An evaluation of the StandFood system was made using a dataset from Slovenia of already classified and described foods using FoodEx2 codes. In the dataset, each food item is represented by a food name and a FoodEx2 code, which is manually added by a human expert. StandFood was then used, first to provide the food category to which the item belongs, and second to describe it using FoodEx2 code. This was then compared with the food category and the code that were added manually.

Classification part Classification is a supervised ML approach that trains a model using a labeled training set to predict a category (class) from input features [21]. The training set consists of instances (observations) described by features whose category membership is known. An algorithm used for classification is called a classifier and it maps input data to a category. The main goal of such an algorithm is to analyze new unseen instances that are not present in the training set, or to predict their categories. Because the ML supervised algorithms perform well for instances from the training set, their evaluation needs to be performed using a test set that consists of instances not found in the training set. For this purpose, the training set is often randomly split into three portions, the training set, the validation set, and the test set. The training set is used to train a ML supervised algorithm, the validation set is used to select the best performing algorithm (or parameters of the algorithm), and the test set is used to evaluate the performance. Some literature suggests that a typical ratio to split into training, validation, and test sets needs to be made using the 60/20/20% or 70/10/20% rule [21]. Another approach for evaluating the performance of ML supervised algorithms is to use k-fold cross-validation [21]. In this case, the training set is divided into k subsets. Each time, one of the k subsets is used as the test set and the other $k-1$ subsets are combined to form a training set. Then the average error across all k trials is computed. The advantage of this method is that it is not important how the data is divided. Every data instance gets to be in a test set exactly once, and gets to be in a training set $k-1$ times. The variance of the resulting estimate is reduced as k increases. The disadvantage of this method is that the training algorithm has to start from scratch k times, which means it takes k times as much computation to make an evaluation. In the case of k-fold cross validation, the test sets are validation sets that are used to select the best algorithm (or best parameters of the algorithm). Then, the selected model needs to be evaluated on a test set that consists of new unseen instances.

When working with ML supervised algorithms, ensemble learning is used [210], which is the process by which multiple classifiers are combined to solve a particular problem in order to improve the performance that can be achieved when each of the combined classifiers are used alone.

The StandFood classification part consists of the following three steps

- Pre-processing of the instances (food items names)
- Feature selection (building a document-term matrix and adding more relevant features)
- Model training

In the case of StandFood, the training set consists of 5,416 instances. Each instance is represented by a food item name and the category to which it belongs (r, d, s, or c). In this case, it involves working with text data, which is unstructured and depends on how people express themselves. Before training a model for classification, the data needs to be pre-processed. First, each food item name is pre-processed by changing all letters to lower case, removing punctuations and numbers. Then, for each food item name, tokenization is applied, which is a process of breaking text into words that are called tokens. Each token consists of a string of characters without a space. For example, one instance is "dried vine fruits (raisins etc.)" and its food category is d. After pre-processing, this instance is transformed into "dried vine fruits raisins etc". Then, tokenization is applied, and the tokens are: dried, vine, fruits, raisins, and etc. These tokens can then be analyzed by applying lemmatization [67] or stemming [68], [69]. For example, the lemma for "dried" is "dry", while the stem is "dri". In the case of StandFood, stemming is applied to each token. After pre-processing, each of the instances represents a document in a text corpus.

The corpus is then transformed into a document-term matrix i.e. feature vectors format. A document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. The matrix gives the relationship between terms and documents, where each row stands for a document (food item name) and each column for a term, and an entry is the number of occurrences of the term in the document. For example, the food item "dried vine fruits (raisins etc.)" is a document, which is a row in the document-term matrix. This document consists of five terms, "dri", "vine", "fruit", "raisin", and "etc", which are the stems of the tokens from the food item name. So, there is an entry 1 on the positions where the raw is "dried vine fruits (raisins etc.)" and the columns are "dri", "vine", "fruit", "raisin", and "etc", because each of these terms are presented exactly once in the food item name. As in the most cases when working with text data the matrix is very sparse, in our case the sparsity is equal to 99%. For StandFood, only the non-sparse entries are used, which means that the terms that only appear in at most 1% of the documents are removed. The document-term matrix consists of 5,416 documents each of which is represented by 276 terms (features). Because the performance of a ML algorithm depends also on the features used to represent the instance, a feature selection is often applied, which is a process of selecting a subset of relevant features to be used to construct a model [21]. In the case of StandFood, removing the sparse terms is a kind of feature selection. However, to improve the model further, four additional features were added. The first is the number of nouns found in the food item name. This is relevant because nouns carry the most information, so more nouns means a higher probability that the food item is a composite food. The second and the third are the number of adjectives and verbs, respectively. Adjectives explain food items in the most specific form (e.g., frozen, fresh) and the verbs are generally related with the method of preparation (e.g., cooked, drained).

This is also relevant because if food items consist of adjectives and verbs, there is a higher probability that it is a derivative or composite food since cooking changes the nature of the food. The last feature is the length of the description. A long description usually means a higher probability that the food item is a derivative or composite food. After representing each instance of the training set by the selected features, the next step is to train the model. For this purpose different classification algorithms can be used. In the case of StandFood, ensemble learning is applied, which combines four classification algorithms: Support Vector Machine (SVM) [211], Random Forest (RF) [212], Boosting (Boosting) [213], and Max Entropy (Maxent) [82]. Using an ensemble of these four algorithms, the model is trained and can be used to predict the category of new unseen instances. A flowchart of the StandFood classification part is presented in Figure 5.1.

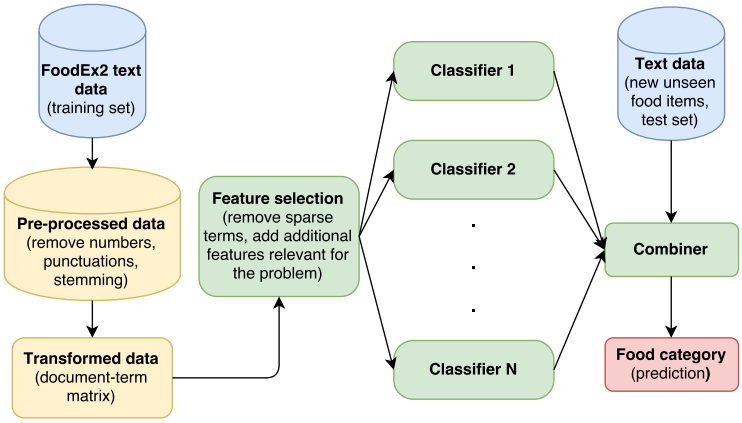


Fig 5.1: StandFood classification part flowchart.

Description part After clarifying the category of the food item, it is then necessary to describe it using the FoodEx2 code. The StandFood description part uses a NLP approach combined with probability theory in order to match a food item using its name to a food item that already exists in the FoodEx2. A similar approach was used to match Internet recipe ingredients with food composition database in Eftimov et al. [33]. However, in the case of StandFood, this approach has been extended.

In StandFood, each food item is described by its name. To describe it using FoodEx2 code, Part-Of-Speech (POS) tagging is used to identify nouns, adjectives, and verbs. The FoodEx2 food names are then searched for all extracted nouns from the POS tagging and each name that consists of at least one of the provided nouns is returned. The result is a subset of food items names from the FoodEx2 data. Then, for each food item name in the subset, POS tagging is used to extract the nouns, adjectives, and verbs. For each set lemmatization is applied to obtain the lemmas.

For example, if the food item, "dried vine fruits (currants, raisins and sultanas)" needs to be described with a FoodEx2 code, first POS tagging is applied to extract the nouns, adjectives, and verbs. These sets are then processed by applying lemmatization. The resultant sets are the noun set, which is {vine, fruit, currant, raisin, sultana}, the adjective set, which is empty for this food item, and the verb set, which is {dry}. The 5,416 FoodEx2 food names are searched for all extracted nouns, which are in the noun set, in this case for the five nouns, and the food names that consist of at least one of the five extracted nouns are returned. For this food item, 225 FoodEx2 food names are returned back. Some of them are: "fruit soup dry", "african oil palm fruits", "native currant", "dried vine fruits (raisins etc.)", "malaga raisins flavour", "juice red currant", "sultanas flavour", etc. Each one of the

returned FoodEx2 food names is processed by applying POS tagging and lemmatization. If the food item is "mushroom soup", after POS tagging and lemmatization, the extracted noun set is {mushroom, soup}, while the adjective and the verb sets are empty. The 5,416 FoodEx2 food names are searched for the two nouns, and the food names that consist of at least one of the two extracted nouns are returned. For this food item, 65 FoodEx2 food names are returned back. Some of them are: "field mushroom", "canned mushrooms", "mushroom flavour", "mushroom soup", "mushroom salad", "pizza and similar with cheese and mushrooms", "honey mushrooms", etc. Each one of the returned FoodEx2 food names is processed by applying POS tagging and lemmatization.

The next step is to define the similarities between the food item name and each of the FoodEx2 food item names that belong to the subset. To do this, the probability of the similarity event calculated using Equation 5.7 is the weight assigned to each matching pair. Finally, the pair with the highest weight is the most relevant found match.

In the example of "dried vine fruits (currants, raisins and sultanas)", the probability of similarity between the food item and each one of the 225 returned FoodEx2 food names must be defined using Equation 5.7. Finally, the matching pair with the highest weight is the most relevant one. In this example, the most relevant match is "dried vine fruits (raisins etc.)" and its probability weight is 0.4. In the example of "mushroom soup", the probability of similarity between the food item and each one of the 65 returned FoodEx2 food names must be defined. The matching pair with the highest probability is "mushroom soup", which is a perfect match, and its probability is 0.5. The flowchart of the StandFood description part is presented in Figure 5.2.

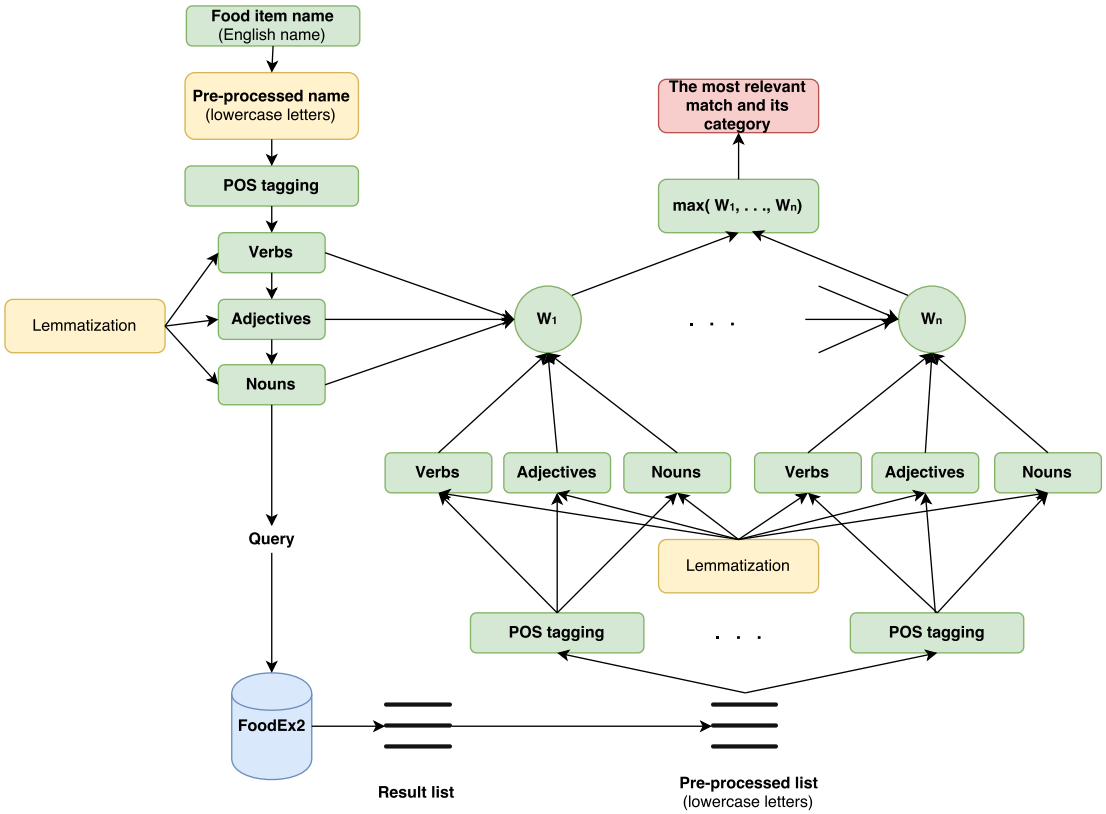


Fig 5.2: StandFood description part flowchart.

Post-processing rules Results of the description part are used to improve the performance of the classification part. Because for each food item the most relevant item from FoodEx2 is returned, the category of the returned food item can be used for post-processing rules. To do this, StandFood applies the following four rules:

- In the first rule, processes that change the nature of a food product are used. Their definitions are in the technical report of FoodEx2 and include, for example, canning, smoking, frying, and baking. Then, for each process in the list of processes lemmatization is applied to avoid different word forms of food items names. So, if a food item is classified as raw (r) using the StandFood classification model but its set of adjectives and verbs (their lemmas) consists of at least one cooking process that changes its nature, it is automatically changed to a derivative (d).
- In the second rule, if a food item is classified as being either raw (r) or a derivative (d) and the result from the description part of StandFood is that the most relevant food item is a composite food (s) or (c), it is automatically changed to a composite food.
- In the third rule, if a food item is classified as a simple composite food (s) and the result from the description part of the StandFood system is that the most relevant food item is an aggregated composite food (c), it is automatically changed to an aggregated composite food (c).
- The fourth rule works in reverse, by changing an aggregated composite food (c) to a simple composite food (s).

The flowchart showing the StandFood post-processing part is presented in Figure 5.3.

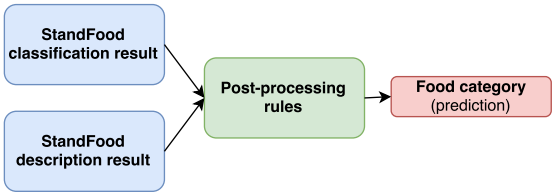


Fig 5.3: StandFood post-processing part.

Results

For the classification part of StandFood, the RTextTools package for the R programming language was used. More details about how to use this package are available in "RTextTools: A supervised learning package for text classification" [214]. For the description part and for extracting information about additional features used in the classification a coreNLP package for R programming language was used [24].

StandFood classification results The training set of 5,416 instances is described using 280 features, from which the first 276 are the non-sparse terms selected from the document-term matrix and the last four are the additional features, which are the number of nouns, adjectives, and verbs, and the length of the food item's name. The problem remains how to classify instances into four categories (r, d, s, and c), which makes it a multiclass classification problem. The distribution of the categories in the training set is: r (2558), d (1795), c (309), and s (754). To this data, different classification algorithms are applied.

To evaluate the algorithmic performance, it is possible to analyze performance metrics such as precision, recall, and accuracy. For classification tasks, the terms true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), were used to compare the results of the classifier with results provided by a human expert. The terms positive and negative refer to the classifier prediction, while the terms true and false refer to the prediction made by the human expert. Using these terms the precision, recall, and accuracy of the classifier are defined as

$$\begin{aligned} Precision &= \frac{TP}{TP+FP}, \\ Recall &= \frac{TP}{TP+FN}, \\ Accuracy &= \frac{TP+TN}{TP+TN+FP+FN}. \end{aligned} \tag{5.9}$$

Precision for a category is the number of true positives (the number of items correctly labeled as belonging to the category) divided by the total number of items labeled as belonging to the category (the sum of true positives and false positives, which are items incorrectly labeled as belonging to the category). Recall is defined as the number of true positives divided by the total number of items that actually belong to the category (the sum of true positives and false negatives, which are items which were not labeled as belonging to the category but should have been). In a classification task, a precision score of 1.00 for a category C means that every item labeled as belonging to category C does indeed belong to category C (but says nothing about the number of items from category C that were not labeled correctly) whereas a recall of 1.00 means that every item from category C was labeled as belonging to category C (but says nothing about how many other items were incorrectly also labeled as belonging to category C). Usually, precision and recall are not discussed separately. Accuracy is the ratio of the corrected predictions and the total number of predictions made.

In the case of StandFood, different classification algorithms were applied, and the obtained models were evaluated using a 10-fold cross validation. The algorithms used were: support vector machine (SVM) [211], scaled linear discriminant analysis (SLDA) [215], random forest (RF) [212], maximum entropy (Maxent) [82], boosting (Boosting) [213], bagging (Bagging) [216], classification tree (TREE) [217], and neural networks (NNET) [218]. Different classification algorithms were used to select the best classification models that will be further used in an ensemble learning. The selected algorithms are the most commonly used algorithms for the classification task described in the ML literature and the accuracy of each of the models using a 10-fold cross validation is presented in Table 5.2.

Table 5.2: Classification accuracy for each ML algorithm using 10-fold cross validation.

	SVM	**SLDA**	**RF**	**Maxent**	**Boosting**	**Bagging**	**Tree**	**NNET**
Accuracy (%)	88.50	72.41	88.95	89.21	85.88	83.47	69.02	77.12

From the results, it becomes clear that SVM, RF, Maxent, Boosting, and Bagging outperform the other algorithms, because from a practical point of view they have better accuracy. To enhance the accuracy even further, an ensemble learning was used. Ensemble learning refers to whether multiple algorithms make the same prediction concerning the category of an instance. For the StandFood classification part, ensemble of the four best performing algorithms: SVM, RF, Maxent, and Boosting, using a majority vote strategy was used [219]. The majority vote strategy uses the category provided by each algorithm

(classifier) and then chooses the category that receives the largest total vote. The obtained accuracy of the ensemble using a 10-fold cross validation is 91.20%. Also, different ensembles combining three or five of the best performing algorithms were used, but the results are not practically significantly different from the ensemble that combines the four best performed algorithms. In the context of scientific research, it is instructive to make a distinction between statistical significance and practical significance. For example, the obtained accuracy of the ensemble of the five best performing algorithms using a 10-fold cross validation is 90.50%. If the results for both ensembles, using the four and five best performing algorithms, respectively, are compared, between them statistical significance can be found, but in practical sense, this difference is not significant. Because of this, StandFood uses an ensemble that combines the four best performing algorithms. The evaluation of the trained model using a 10-fold cross validation gave the most promising results and the model can be used to classify new unseen instances. To evaluate StandFood properly, 532 new instances from food composition database were used. The results for precision and recall are presented for each category separately in Table 5.3.

Table 5.3: Precision and recall for each food category using the evaluation set.

Category	**Precision**	**Recall**
r	0.72	0.99
d	0.81	0.81
c	0.75	0.67
s	0.95	0.57

The recall for the raw category (r) is 0.99, which means that 99% of the raw instances in the test set are assigned to the raw category. The precision for the raw category (r) is 0.72, from which it follows that there are many false instances assigned to this category. For the derivative category (d) the precision and recall are 0.81 and 0.81, respectively. For the aggregated composite food category (c), the precision is 0.75, which indicates that there are false instances also assigned to this category. Analyzing the false instances that appear in this case showed that these instances belong to the simple food category (s) and the StandFood classification part assigned them as aggregated composite food (c). The recall for the aggregated food category (c) is 0.67, which indicates that only 67% of the aggregated composite food instances presented in the test set are assigned to that category. Looking at the results, the majority of improperly classified instances from the category are assigned to the raw category (r), which influences the precision of the raw category (r). The precision of the simple composite food category (s) is 0.95, which means that there are only a few false instances assigned to that category, while the recall is 0.57, which indicates that only 57% of the simple composite food instances from the test set are assigned to that category. In most cases, the improperly classified instances from that category are assigned to the raw (r) or derivative (d) categories.

Table 5.4 gives the results from the StandFood classification part of eight randomly selected but correctly classified instances, two per food category.

StandFood description results In the description part, the main goal is to describe food items according to FoodEx2 and to find the most relevant food item that exists in the FoodEx2 data. As well as the FoodEx2 code, the food category (r, d, s, and c) is returned, which is further combined with the result from the classification part to improve its performance.

The StandFood description part is an NLP approach combined with a probability

Table 5.4: Correctly classified instances by the StandFood classification part.

Food item	**Category**
Barley grains	r
Mandarins (citrus reticulata)	r
Buckwheat flour	d
Oat flakes	d
Fruit compote	s
Marmalade, mixed fruit	s
Rice and vegetables meal	c
Mushroom soup	c

theory. For its evaluation, for each food item from the test set, it returns the most relevant food item from FoodEx2. Then, the obtained FoodEx2 code is compared with the FoodEx2 code that was manually assigned to the food item by a human expert. Evaluating the description part on the test set, 79% of the instances obtain relevant FoodEx2 codes. In the set of instances that obtain relevant FoodEx2 code, there are instances for which FoodEx2 code assigned by StandFood is the same as FoodEx2 code provided by a human expert. There are also instances in which FoodEx2 codes are different but are close together in the FoodEx2 hierarchy, so they are still related to the same food. For example, for the food item "mutton / lamb meat (ovis aries)", the FoodEx2 code assigned by StandFood is A01RK, which is related to the food item "Lamb fresh meat", while the FoodEx2 code provided by the human expert is A01RH, which is related to the food item "Sheep muscle". In this case, both codes are very close in the FoodEx2 hierarchy, and A01RH is a parent for A01RK. And there are instances for which the manually assigned code is incorrect, while the StandFood system returns the correct FoodEx2 code, which can happen because a human expert is not familiar enough with the details in FoodEx2. In addition, StandFood can return more than one relevant choice, when more food items have the same highest weight. In such cases, the user should select which of them is the most relevant one. As an example, ten randomly selected instances are presented in detail in Table 5.5.

For the first seven food items, the FoodEx2 code obtained by StandFood is the same as the FoodEx2 code manually assigned. Having the true FoodEx2 code, more information for a food item can be obtained by the FoodEx2 data with a simple search using the FoodEx2 code. For example, "mushroom soup" is described as A041R#F02.A06GY$F04.A0ETG$ F28.A07MR$F28.A0BA1. This means that the product is "mushroom soup", with the part-nature descriptor (F02) "soups ready-to-eat (as part-nature)" (A06GY), then it has an ingredient descriptor (F04) set to "fungi" (A0ETG) and two process descriptors (F28), one for "reconstitution from concentrate, powder or other dehydrated form" (A07MR) and the other for "cooking and similar thermal preparation processes" (A0BA1). For the eighth food item, the FoodEx2 code obtained by StandFood and manually code differs. For StandFood, it is a perfect match whereas for manually assigned FoodEx2 code, food item is "wheat semolina" (A004F). The two codes are close in the FoodEx2 hierarchy. However, the StandFood code is more relevant since the code is for "wheat flour durum" that exists in FoodEx2. For the ninth food item, StandFood again returns a perfect match. The manually assigned FoodEx2 code for this food item is "gingerbread dough" (A009Q) and has a final-preparation descriptor (F14) set to "baking" (A07GX). However, "gingerbread" already exists in FoodEx2 data and there is no need to recode it (FoodEx2 recommends to code new food items only if they do not exist in the FoodEx2 data). Also, if there is a need to code new food items, the FoodEx2 recommends not to use facet descriptors from F13

Table 5.5: Result from the StandFood description part for ten randomly selected instances.

Food item	**StandFood FoodEx2 code**	**StandFood relevant FoodEx2 item**	**Manual FoodEx2 code**
Mushroom soup	A041R	Mushroom soup	A041R
Prepared green salad	A042C	Mixed green salad	A042C
Meat burger	A03XF	Meat burger no sandwich	A03XF
Yeast	A049A	Baking yeast	A049A
Brown sauce (gravy, lyonnais sauce)	A043Z	Continental European brown cooked sauce gravy	A043Z
Cow milk, < 1% fat (skimmed milk)	A02MA	Cow milk skimmed low fat	A02MA
Supplements containing special fatty acids (e.g. omega-3, essential fatty acids	A03SX	Formulations containing special fatty acids e.g. omega3 essential fatty acids	A03SX
Durum wheat flour (semola)	A004C	Wheat flour durum	A004F
Gingerbread	A00CT	Gingerbread	A009Q$F14.A07GX
Cherry, fresh	A01GG	Cherries and similar	A01GK
	A01GH	Sour cherries	
	A01GK	Cherries sweet	
	A0DVN	Nanking cherries	
	A0DVP	Cornelian cherries	
	A0DVR	Black cherries	

to F16 (F13- cooking method, F14- final preparation, F15- preservation technique, F16 - structural treatment), and instead use only F28 i.e. the process descriptor. In this case, the code given by StandFood is more relevant than the manually assigned code. In the last example, the manually assigned FoodEx2 code is one of the codes returned by StandFood. In this case, there are several food items from the FoodEx2 data that have the same weight, which is the highest weight obtained by the equation used in the description part, so they are all returned as relevant matches. The user should select which of the returned matches is the most relevant one. In the case of StandFood, 10% of the food items that received a relevant FoodEx2 code have a more relevant code than the code assigned manually which can happen because a human expert is not familiar with the details in FoodEx2.

The idea used in the description part of StandFood is based on NLP, or using text similarity measures. StandFood uses short segments of texts, so it can not use standard text similarity measures [107] because they fail when tasked with computing the similarity between two very short segments of text [108]. In this work, different similarity measures, including the Jaccard index applied on the character level between the names of two food items, Jaccard index applied on the word level, negative Kullback-Leibler (KL) divergence [220] were used, but none gave promising results. The problem of using short segments of text is explained in [108]. The StandFood description part gives good results because domain modelling is applied, and the model combines text similarity measures using probability theory.

StandFood post-processing rules After obtaining the result from the StandFood description part, the food category (r, d, s, or c) of the most relevant match is further combined with the obtained food category from the StandFood classification part. By applying the four post-processing rules defined in Section 5.4.3, the newly obtained precision and recall for each food category using the test set are obtained in Table 5.6.

A comparison between Table 5.3 and Table 5.6 reveals improvements in the precision

Table 5.6: Precision and recall for each food category after applying the post-processing rules.

Category	Precision	Recall
r	0.85	0.99
d	0.90	0.84
c	0.82	0.87
s	0.97	0.83

and recall of each category and 89.28% of the instances from the test set are correctly classified. As an example post-processing of seven randomly selected foods were carried out (Table 5.7).

Table 5.7: Food categories for seven food items after post-processing rules.

Food item	Classification part category	Post-processing category
Cabbage Chinese boiled	r	d
Marzipan	r	s
Gingerbread	r	c
Water, bottled, flavoured, citrus	d	s
Salad, tuna-vegetable, canned	d	c
Multigrain rolls	c	s
Croissant, filled with jam	s	c

Discussion on StandFood

The main benefit of using StandFood is that missing FoodEx2 data in a lot of food composition data sets can be obtained. Being a semi-automatic system, it significantly reduces the time needed to code food items. When the FoodEx2 data is available, it is easier to compare the food data presented in the data set with food data from another study, or combine them. StandFood gives 79% correctly classified and described instances, which means that it is a promising tool that can be used. However, 21% of the instances are not correctly described (in them there are instances that are correctly classified), and this happens according to the fact that food items do not exist in FoodEx2, food items are typical for some cultures, or food description is not complete and detailed as possible. StandFood is a semi-automatic system, which means that the user needs to check the result that is obtained. Also, it is seen that there are some cases when the StandFood system gives more relevant FoodEx2 code than a human expert. This happens because human experts manually code food items with facet descriptors to describe them in detail, when food items already exist in FoodEx2, and they are not familiar with the details in FoodEx2, which makes this an ongoing problem, as FoodEx2 will evolve with new entries all the time. When a new food item is not present in FoodEx2, it needs to be coded (in most cases they are composite foods), which will be our future work. In some cases, StandFood returns the ingredients of the foods that are presented in the food description that further need to be coded with some other facet descriptors. For example, for the food item "baked potatoes with parsley", it returns "potatoes" and "parsley". To code the new food items (composite foods), it is better to find recipes for them and try to extract all useful information that can help the process of coding, starting from the used ingredients, coking processes, etc.

For our future work, it is planned to code these cases following the recommendations given in the FoodEx2 technical report.

Using the trained classification model on the FoodEx2 data, there are some weaknesses when it is applied on new instances. The main reason why this happens is because of the intrinsic ambiguity of language, meaning that different people may interpret the name of a food item differently. For example the instance "potato boiled" is a derivative food, however the StandFood system assigns it to the raw category. This happens because "boiled" is not used as a feature for classification and it is one of the sparse terms that are removed because it appears only in 3 out of the 5,416 instances. Using synonyms is also problematic (e.g., mushrooms tinned (r), mushrooms canned (d)), and this happens because "canned" exists as a feature extracted from the FoodEx2 data. Despite the weakness of the trained model, it can still be used because the StandFood classification part is improved using post-processing rules.

According to the time needed to code the new 532 food items, in the case of StandFood five minutes are required because the current version of StandFood is programmed in a sequential way. Parallel programming would allow the same task to be carried in seconds. A human expert takes on average 10 minutes per item, or 5,320 minutes for 532 food items, or 11 days if a human expert works 8 hours per day. This information is provided by a human expert who has experience working in food matching. However, the result from StandFood needs to be checked and in the case of several options to choose the correct one. For 532 food items, through checking and choosing the correct result when several options were given took one working day (8 hours).

StandFood depends on the food item name. Since different people may interpret the name of a food item in a slightly different way, a better explanation in the food item name would provide a higher probability of finding the correct FoodEx2 code. At present, the system works using English food item names. In our case, the names of food items were translated from Slovenian by a non-native English speaker, which also represents a potential weakness because the description might include errors or strange translations. In future work, it is also planned to upgrade StandFood so it can be used with different languages, by automatically translating food items names into English using the Google Translation API. The Google translate API is not perfect, especially when local food items names can be very peculiar and difficult to translate. Translation will work better when the food item name is as complete and detailed as possible. So we will try to find a way how to obtain more detailed and complete food item names that will be used in translation and then by StandFood. Also, we will try to link other food classification systems (i.e. LanguaL) to FoodEx2 in order to improve the result of StandFood.

In practice, StandFood can be used to find missing FoodEx2 codes for food composition data and food consumption data that are basic resources that need to be linked and combined for dietary assessment methods. For the near future, we are planning to implement StandFood as a web service, which will be available for the wider research community.

5.5 Discussion and Conclusion on String Similarity of Domain-Specific Short Text Segments

String similarity measures are the basis of text normalization methods. The second part of our second hypothesis is that by probability modeling of the domain using the morphological information, string similarity measures of domain-specific short segments of text can be improved. The hypothesis is confirmed on the food matching problem, where the newly proposed method shows a very promising performance over standard string similarity measures.

Our method uses POS tagging to obtain the morphological information presented in the text. Then, it combines the relevant morphological information for the domain of interest by using the probability theory to model the domain [33]–[38]. Compared to other text similarity measures (e.g., Cosine, Jaccard, Jaro, Levenshtein, etc.) that can fail in string similarity between two very short text segments, the proposed method gives promising results considering a real problem in food science, food matching.

The benefit of using the proposed method is that it can easily be extended for each domain of interest. The result of it can be further combined with some other standard string similarity measures using some heuristics in order to achieve better matches. Also, it can be used as part of other systems for ranking candidate matches and more further analysis. For 532 short segments of text, the time complexity of the POS tagging-probability weighted method is 30.49 s, if the pre-processing and the POS tagging of the text segments that belong to the terminological resource, which is searched for the 532 segments, is made in advance and the result is stored as resource data for the method.

The limitation of the method is that the relevant morphological information for a specific domain needs to be selected by a human expert. However, working on a concrete application it is not a problem and it is not a time consuming task. For our future work, we are also planning to find a way of automatically detecting the relevant morphological information and presenting a more general version of the method. Also, we would like to extend this method using distributional semantics methods.

Comparing text strings is a common and fundamental task in many statistical text-processing applications. The goal of string similarity measures is to quantify the similarity between two text strings. They are the basis for text normalization methods, which are the processes of automatic mapping between a concept in text and a concept in a terminological resource that is used for the domain. The text normalization methods are crucial when working with information extraction in a given domain because in most cases the extracted entities are presented as a short segment of text (phrases). The problem that appears is that the same entity can be represented in different ways regarding the text variability (the way of how people express themselves, their writing style, etc.), so applying text normalization methods helps collect and combine the information for a given entity. For this reasons, we proposed a novel method for string similarity of domain-specific short segments of text, called POS tagging probability weighted method. The method uses POS tagging to obtain the morphological information presented in the text. Then, the probability theory is used to model the domain using the relevant morphological information that describes the domain. Experimental results obtained in the food domain, especially the food matching problem, have shown that the method gives the most promising results compared with some standard string similarity measures.

Chapter 6

Discussion of a Real-World Application in Nutrition Science

To see the contribution of each of the proposed methods to the nutrition science, in this chapter, we provide a discussion of a possible real-world application that can benefit of using them.

6.1 Discussion on DSC

Over the last decades, there have been revolutionary developments in nutrition science, which is a consequence of availability of advanced methodologies for analysis, as well as assess to bioinformatics databases. Currently, statistics has become an increasingly important tool to quantitatively analyze information and is a part almost of each scientific paper published in nutrition science journal [221]. In most cases, papers include reporting confidence intervals and results obtained of making a statistical comparison. Statistical comparison has a huge application in nutrition science and can be used to study food behavior regarding for example different age or gender groups (different populations), to study patterns in food choice (e.g., which foods are selected in combination? etc.), modelling long-term intake using short-term measurements, or food consumption surveys, where the main purpose is to identify, predict and detect chronic or acute food and nutrition problems, to plan and implement food and nutrition programmes and interventions, and to monitor changes and evaluate the impact of interventions and programmes. The problem is in selecting the right statistic to apply as a specific measure. For example, researchers often report either the average or median without being aware that averaging is sensitive to outliers and both, the average and median, are sensitive to statistical insignificant differences in the data. Even reporting the standard deviation of the average needs to be made with care since large variances result from the presence of outliers.

To present the difference that can exist between the common approach for making a statistical comparison and our proposed approach, we show an illustrative example by analyzing data from Slovenian national survey on food consumption, which was initiated by the National Institute of Public Health in Slovenia. We analyze results from 200 respondents of a food frequency questionnaire, FFQ. A FFQ consists of food groups that includes foods and beverage and the main idea is to obtain frequency, or portion size, over a specified period of time. In our case, we are interested if there is a difference between different age groups regarding a specified food group for a population and not an individual. The 200 respondents were split in three groups (e.g., populations) following the recommendations from nutrition scientists:

- Group 1: $18 \leq \text{Age} < 25$;
- Group 2: $25 \leq \text{Age} < 51$;
- Group 3: $51 \leq \text{Age} \leq 65$.

To compare these three groups on a specified food group, we focus on a single question from the FFQ, which is *"How often do you enjoy the following foods from the Grain and Cereal products group?"*. This food group, F, consists of 14 foods, f_i, $i = 1, \ldots, 14$, on which each respondent can give one of six possible answers:

- 1: Once or twice per day or more;
- 2: Four to six times per week;
- 3: Twice or three times per week;
- 4: Once per week;
- 5: Once to three times per month;
- 6: Never.

For the comparison, we present the result using the common approach and using the proposed DSC approach. A single question from the FFQ regarding a food group corresponds to a multiple-problem analysis, while the analysis performed on a single food that belongs to that food group corresponds to a single-problem analysis. Following the common approach for making a statistical comparison, the answers obtained from each group on a single food are averaged and the value is set as a representative value for the multiple problem analysis. However, averages are sensitive to outliers and small insignificant differences (ϵ-neighborhood) that can exist between compared data values, which may lead to incorrect conclusions. For this reason, the DSC approach is also used which is based on comparing distributions instead of using only one statistic as a representative value. The DSC approach that was presented in Chapter 3 works with continuous data, while in this use case the answers are categorical ordinal data. We can still use the same approach, but from a statistical point of view it will not be correct because the two-sample KS test and the two-sample AD test can be used only with continuous data.

In order to make DSC work with ordinal categorical data, we use the same idea presented for continuous data, but we only change the criteria for comparing distributions. So, instead of using the two-sample KS or two-sample AD test for comparing distributions, we use a Chi-square two-sample test [222], [223]. The Chi square two-sample test can be used to test if two data samples come from the same distribution. It is based on binned data and the binning should be the same for both data sets. The main idea is that the number of observed data values in each bin should be similar if both data samples come from the same distribution. In the case of the common approach the required conditions for safe use of the parametric tests are not satisfied, so the relevant test is the Friedman test. The ranks obtained by the common approach are presented in Table 6.1a, while the ranks obtained by the DSC approach are presented in Table 6.1b. The p-values obtained by the Friedman test, which is appropriate for both approaches, are 0.00 and 0.46, respectively. Using the common approach we can conclude that there is a significant statistical difference between the three groups, while using the same data with the DSC approach we can conclude that there is no significant statistical difference between the groups.

In the case of the common approach (Table 6.1a), we can see that the groups obtained different ranks on each single food. This happens because the average is used as a representative value and even when the averages are in some small ϵ-neighborhood, they will

Table 6.1: Ranks for population groups.

(a) *Common approach*

F	Group 1	Group 2	Group 3
f_1	1.00	2.00	3.00
f_2	3.00	2.00	1.00
f_3	1.00	2.00	3.00
f_4	1.00	2.00	3.00
f_5	1.00	2.00	3.00
f_6	1.00	2.00	3.00
f_7	3.00	2.00	1.00
f_8	1.00	3.00	2.00
f_9	1.00	3.00	2.00
f_{10}	1.00	2.00	3.00
f_{11}	1.00	2.00	3.00
f_{12}	1.00	2.00	3.00
f_{13}	1.00	2.00	3.00
f_{14}	2.00	1.00	3.00

(b) DSC ranking scheme

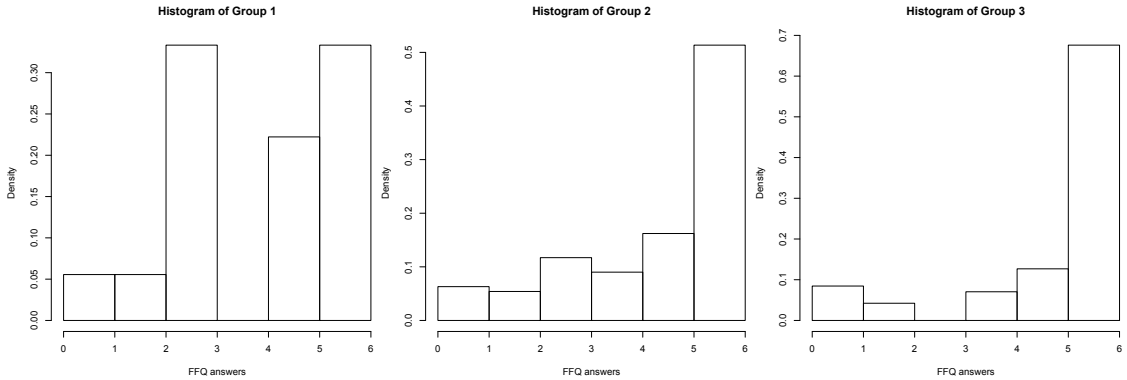
F	Group 1	Group 2	Group 3
f_1	2.00	2.00	2.00
f_2	2.00	2.00	2.00
f_3	1.00	2.00	3.00
f_4	1.00	2.50	2.50
f_5	2.00	2.00	2.00
f_6	1.00	2.00	3.00
f_7	3.00	2.00	1.00
f_8	2.00	2.00	2.00
f_9	2.00	2.00	2.00
f_{10}	2.00	2.00	2.00
f_{11}	1.50	1.50	3.00
f_{12}	1.50	1.50	3.00
f_{13}	2.00	2.00	2.00
f_{14}	2.00	2.00	2.00

be ranked as different. In the case of the DSC approach (Table 6.1b), we can see that groups can also be the same (e.g., f_1, f_2, f_5, etc.) and this happens because the ranking is made using the whole distribution of the data. To see what happens on a single food, in Figure 6.1, we present the histograms of the data for each group separately on three different foods (f_3, f_4, and f_8).

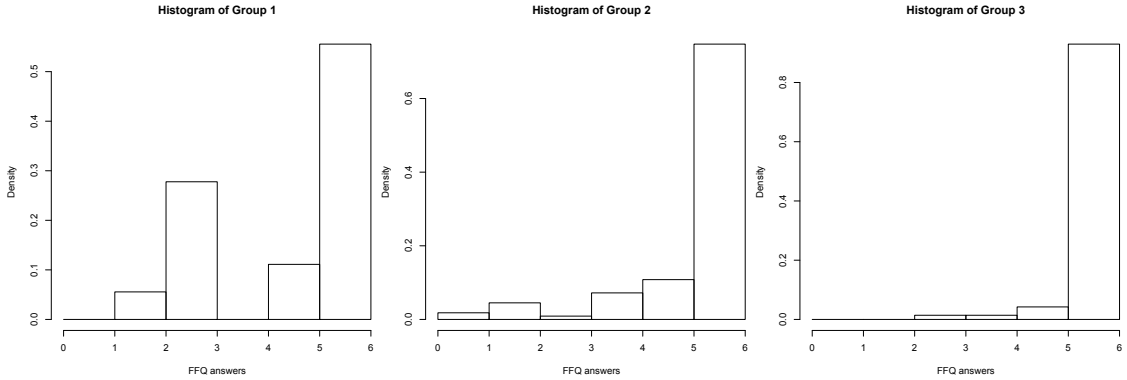
From Figure 6.1, it is obvious that we have three different distributions for f_3, so the groups obtained different DSC ranks. We can also say that the distribution of the first group is different from the distributions of the second and third group and the DSC ranks are 1.00, 2.50, and 2.50, respectively. This is also the result of using the chi square two-sample test. Also, it is obvious that distributions of groups are the same, so the DSC ranks are 2.00, 2.00, and 2.00.

Using the common approach, researchers usually reported and compared the mean or median of the data without variance. Even more, when variance is reported, large variance may often indicate the presence of outliers. In order to be aware of the presence of outliers, the DSC approach is presented and is based on the idea of comparing distributions instead of using only one standard statistic to represent the data. It gives more robust results when the results of the common approach can be affected by the presence of outliers or some small ϵ-neighborhood. Using the DSC result obtained for multiple-problem analysis, we can say that there is no significant statistical difference of how the three groups enjoy the Grain and Cereal products group. However, on a single problem analysis we can see that there can be a difference between the groups. This kind of analyses should be performed and are basis for food behaviour analysis and food choice analysis. Even more, results can be further combined for more advanced analysis such as calculation of dietary nutrient intake or to plan and implement national food and nutrition programmes and interventions.

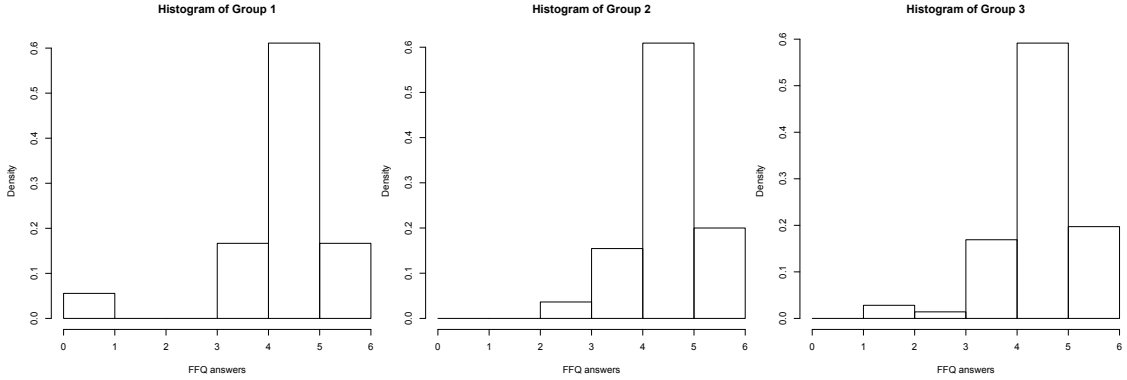
Many researchers have problems and difficulties performing a statistical analysis of their data, which is often crucial in interpreting their results. The problems appear because each statistical test has some conditions (assumptions) about the data that must be satisfied in order to apply the test. Similarly, in nutrition science, as in other research domains, researchers do not check for these conditions and simply apply a statistical method based



(a) Group 1 for f_3 (b) Group 2 for f_3 (c) Group 3 for f_3



(d) Group 1 for f_4 (e) Group 2 for f_4 (f) Group 3 for f_4



(g) Group 1 for f_8 (h) Group 2 for f_8 (i) Group 3 for f_8

Fig 6.1: Histograms of FFQ answers of each group for f_3, f_4, and f_8.

on one used in a similar study. As a result, their conclusions can be incorrect. For this reason, to become familiar with statistical comparison, we provide a short tutorial on how to perform a statistical analysis of food data, presented as a book chapter in the "Science within Food: Up-to-date Advances on Research and Educational Ideas" [39]. In addition, we developed an e-learning tool, which is available at `http://ws.ijs.si/`

`statTool/` (see Figure 6.2). The e-learning tool is based on the common approach for making statistical comparison. It is a Shiny application, which automatically checks the conditions for each method and offers only those methods that are appropriate for the data. Overall, the e-learning tool not only reduces the time needed to perform a statistical analysis but importantly, it can help in interpreting results and increase awareness of using an inappropriate statistical method.

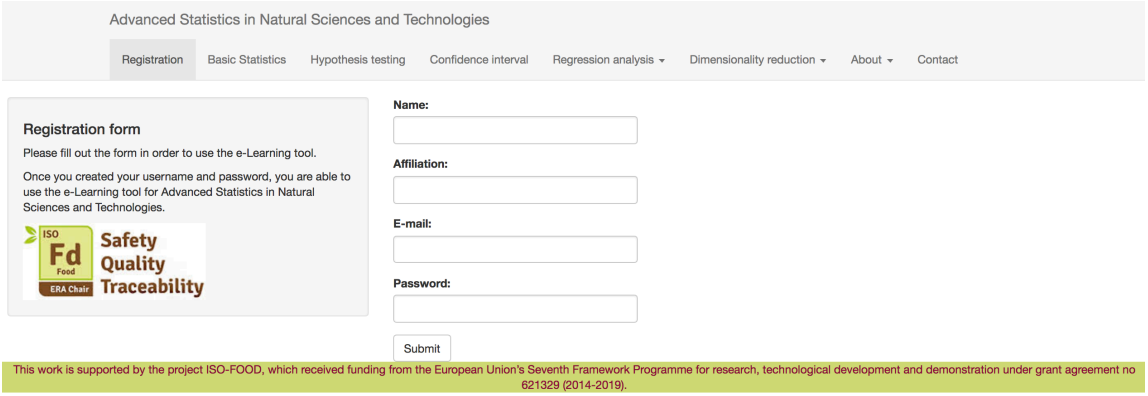


Fig 6.2: E-learning tool: Advanced Statistics in Natural Sciences and Technologies.

The idea for such a tool came about after asking a sample of 28 researchers (master and doctoral level) who work in food analysis to select a method to do the following: i) compare two data samples, ii) compare three or more data samples, and iii) calculate a confidence interval. In the first case, 36.36% selected a *paired-t test*, while 45.46% did not know which statistical test to use. Only 18.18% chose the correct answer, which was to first check the conditions necessary for each method based on their data and then decide on the method to use. In the second case, 81.82% selected to use *ANOVA*, while again only 18.18% chose to check the conditions of each method before selecting which statistical method to use. Finally, when calculating the confidence interval (CI) just over half of the participants (54.54%) selected the standard formula that requires normally distributed data, even though their data is not normally distributed, i.e. they did not know which formula to use. In this case, none of the participants chose to check first the conditions of the data. It is clear that only a small percentage of the participants know how to perform correctly a statistical analysis.

A presentation on statistical analysis was presented to participants of the ISO-FOOD spring school on the use of isotopes in food organized from 4th to 7th April 2017 at Jožef Stefan Institute, Ljubljana, Slovenia. After an introductory presentation, the participants (master and doctoral level researchers) where asked to complete the same three tasks as earlier. After dividing the participants into three groups they were given thirty minutes to perform one of the three examples using a statistical software package in which they were experienced. The data used for the tasks is available at (`http://cs.ijs.si/opendata/DataSets.zip`). Each group then presented their results. The group that worked on task one, began by using Excel, but switched to RStudio [224] since Excel was taking too much time. First, they checked the required conditions for the safe use of a parametric test and because the tests were not satisfied, they selected an appropriate nonparametric test, which in their case was the *Mann-Whitney U test*, whereas before the presentation, the majority of opinion was to use a *paired-t test*. The second group chose to compare the three data samples using three different pairwise comparisons by comparing the variances of the data samples using the *Fisher test* [225]. However, this is not correct because these pairwise

comparisons are independent and combined independent p-values need to be additionally calculated to control the family wise error (FWER), which is the probability of making type I errors when performing multiple hypotheses tests. The group also explained that they performed multiple pairwise comparisons because they could not find the *ANOVA* function in Excel. However, the appropriate statistical test in this example was *Kruskal-Wallis test*. The third group, which worked on calculating CI, also used Excel using the equations for normally distributed data despite this not being the case in the test data. Afterwards, they switched from Excel to SPSS [226], and performed the same experiment reporting the bootstrapping confidence intervals, which were the correct results for this experiment. Prior to the presentation, when asked, the suggestion was to use the sigma, two-sigma, three-sigma rule. Each group was then set the task of reviewing three scientific papers published in a journal with high impact factor and then to judge if the statistical analysis used is correct and what if any information relating to statistics is missing. All groups agreed that in the three reviewed papers the authors only said that they used the *t-test* to compare two data samples and a *one-way ANOVA* to compare more then two data samples. They did not provide the results for checking the required conditions for the safe use of parametric tests, so it is not clear how they selected the statistical test or if they also followed a similar study. In addition, the CIs were reported using the sigma, two-sigma, three-sigma rule without first checking if the data was normally distributed or not. Finally, after a short presentation of the e-learning tool each group repeated each of the three tasks. All three groups needed only a few minutes to perform the three examples. For the first example, they needed only to upload the test data, and to indicate if they have paired or unpaired samples. The e-learning tool checks all other conditions and offers them with the most suitable statistical test, in this case the *Mann-Whitney U test*. For the second example, the tool offered only the *Kruskal-Wallis test*, which meant that they did not have to check for either normality or homoscedasticity of the variances. Finally for the third task, the participants needed to specify only which parameter they want to calculate the CI and the tool provides the CI together with the method used to obtain it.

To provide more robust results to outliers and insignificant statistical differences that can exist between data values, we also developed an e-learning tool for deep statistical comparison, which is available at `http://ws.ijs.si/dsc/` (see Figure 6.3).

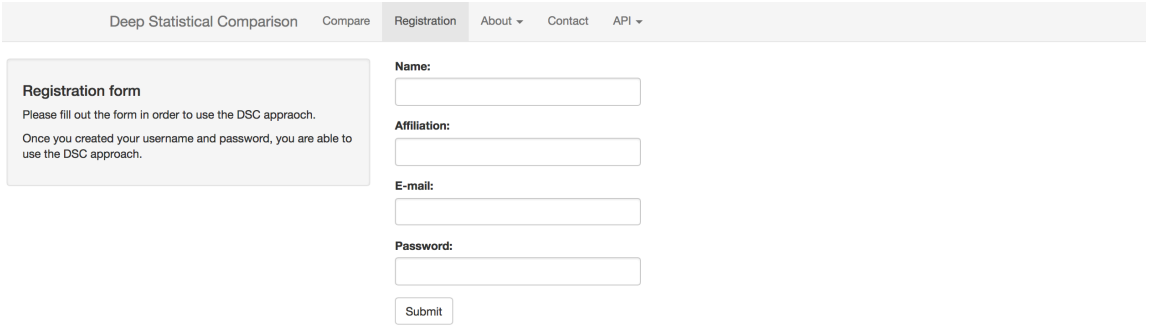


Fig 6.3: E-learning tool: Deep Statistical Comparison.

6.2 Discussion on drNER

In recent years, several European initiatives are focused on research infrastructure for nutrition science. In 2014, the European project "QuaLiFY, Quantify Life - Feed Yoursel"

started (http://qualify-fp7.eu/), where the main goal was to create a solid basis for commercial activities in the area of personalized nutrition by providing a data- and information infrastructure and harmonized open innovation protocols. Also, in mid December 2014, the project "ENPADASI, JPI HDHL European Nutritional Phenotype Assessment Data Sharing Initiative" started (http://www.enpadasi.eu/index.html), where the goal is to deliver an open access research infrastructure that will contain data from a wide variety of nutrition studies. In 2014, the project "Era-Chair ISO-FOOD, Era Chair for isotope techniques in food quality, safety and traceability" started (http://isofood.eu/), in which one goal is the development and updating of a food databases that will be equipped with tools for processing and managing data and information. In Autumn 2015, the project "Richfields" started (https://www.richfields.eu/), where the goal is to design a consumer-data platform to collect and connect, compare and share information about food behaviors, to revolutionize research on every-day choices made across Europe. The common thing for all these projects is that for their development, harmonization is crucial as combining studies depends on mapping those of similar data and design. Making research data available/accessible after study finalization is often a prerequisite for funding. However, large amount of textual data that is a result of a study is also available in scientific publication. So these research infrastructures also need methods for extraction of the relevant information and this is a well-known task in natural language processing (NLP). For this reason, as a part of the "Era-Chair ISO-FOOD" and "Richfields", we developed drNER, which is a rule-based named-entity recognition method for information extraction of evidence-based dietary recommendation and it is focused on food, nutrient, and quantity/unit concepts (entities). This information extraction is also one step that need to be made as a part of semantic enrichment, which adds machine-readable meaning to the extracted concepts and it is important for data sharing, which implies not only collaboration and data exchange, but allow merging of data sets that can be queried as a single integrated data set.

In future, drNER can also be used to produce data sets that will contain the last dietary recommendations according to a specific selected disease and they will be available to clinical dieticians. Currently, it was evaluated using 100 documents, 50 documents are dietary recommendation summary that are extracted from Food and Nutrition Information Center of United States Department of Agriculture and 50 documents that are abstracts of scientific publications. The obtained results are public available and also commented by two clinical dieticians, who agree that promising results are achived.

6.3 Discussion on StandFood

The coding of food composition and food consumption data in the same food description and classification system will enhance data linkage and harmonization, and will reduce errors and time needed for food matching. In 2016, EuroFIR, organized a workshop on food matching at the EuroFIR Food Forum 2016, where different European countries presented their results and experience of matching food items. Our group was a part of that workshop and we presented a method for matching Internet recipe ingredients to food composition data [33], which is a semi-automatic method. The other proposed methods were based on manually matching the food items to a food information terminological resource, which is a time-consuming task. In October 2017, we presented StandFood, which is a semi-automatic system for classifying and describing foods according to FoodEx2, at the third International Conference of Metrology Promoting Standardization and Harmonization in Food and Nutrition [35]. In mid October 2017, food matching was a special topic covered at the 12^{th} International Food Data Conference, at which we presented the results of matching

foods from EuroFIR food composition databases to FoodEx2 by using StandFood [36], [38]. The coded food composition data was from Slovenia, Greece, Italy, and United Kingdom. As a part of the same conference, a FoodEx2 workshop was held, in which was presented how to code food composition data in FoodEx2. After that conference, in November 2017, the Food and Agriculture Organization of the United Nations, FAO, and the World Health Organization, WHO, lunched a FAO/WHO Global Individual Food consumption data Tool (FAO/ WHO GIFT, `http://www.fao.org/gift-individual-food-consumption/en/`), which provides up-to-date inventory of individual quantitative food consumption surveys conducted in low- and middle-income countries. All food consumption data are coded in FoodEx2 and are used to answer questions such as: what are the source of nutrients in people's diet?, what do people eat?, what is the level of consumption among high consumers of different foods?, etc. FAO/WHO GIFT also provides an opportunity to make other food consumptions surveys visible, for which there is an opportunity to use StandFood to find the missing FoodEx2 codes.

The food matching problem is also related to the nutrient intake calculation, which is a problem of combining the information from food consumption data and food composition data. However, this calculation also depends from the quality and the coverage of the data presented in the food composition databases. StandFood can also be used to find what kind of food composition analyses are still missing, which is a time-consuming task when needs to be manually done for each country regarding the property of biodiversity. To show this, we present an illustrative example in which we try to find which of the frequently used recipes' ingredients together with their analyses are represented in the EuroFIR food composition database [6]. As our target data, we selected recipes in English. First, we collected data from a free Internet data source (`http://allrecipes.com/`). The data we used is a collection of 1,000 recipes written in English and we collected it using HTML parser, written in the *R programming language*, to extract information from a free recipes web site. In Figure 6.4, we present a word cloud of the recipes ingredients.

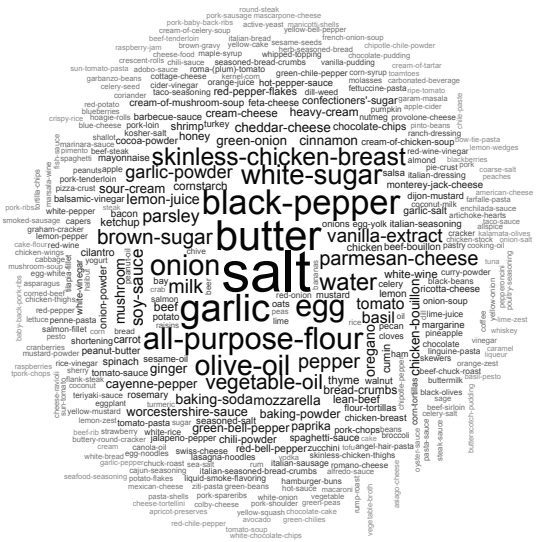


Fig 6.4: Word cloud of recipes ingredients.

Further, we found which of these ingredients are represented in the EuroFIR database. For this purpose we selected the databases from UK, Denmark and Slovenia. We compared the name of each ingredient found in the recipes with the English name attribute of the

food entity from the database, for which StandFood was used to perform food matching. Finally, we found that 65.33% of the ingredients are represented in the EuroFIR database and 34.67% are missing. In Figure 6.5, we present a word cloud of the ingredients for which we found analyses in the EuroFIR database. This information is valuable for data compilers, who compile compositional data for food items and need to know which data is required but is still missing. To understand the complexity of this kind of work, let us explain that in a food composition database an average number of food items is between 5,000 and 10,000 and each food item may have up to 270 parameters. In Europe, each country has its own food composition database.

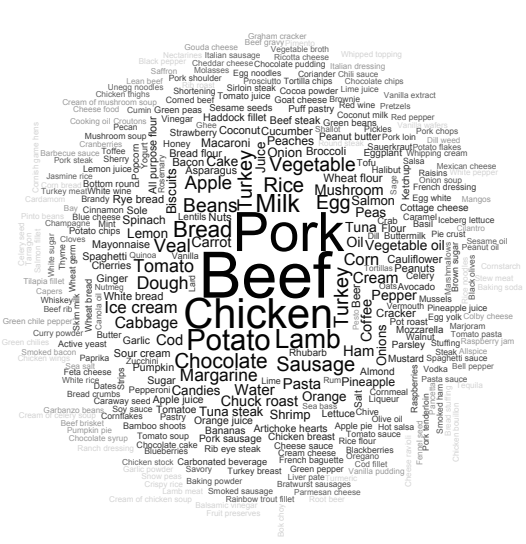
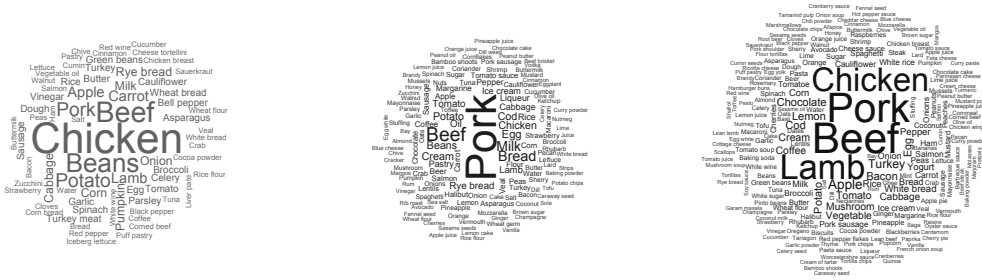


Fig 6.5: Word cloud of ingredients for which we found analyses in the EuroFIR database.

Also, we analyzed the databases from UK, Denmark and Slovenia, separately. In Figure 6.6, we present word clouds of ingredients for which we found analyses in the UK database, Denmark database and Slovenian database, respectively.

A part of StandFood was also used for harmonization of food- and nutrition-related data as a part of the European project "QuaLiFy, Quantify Live - Feed Yourself" [227], where the main goal was to integrate the results of FP6 and 7 projects EuroFIR, NuGO and Eurreca, Eurogene and Food4Me. By using POS tagging-probability weighted method we defined similarity matrix that can be used to find similar terms described on different ways. Using this method we introduced an ontology that can be used in the domain of food-related data and it covers wider domain instead of recombining several existing ontologies, which are developed for very specific uses. The ontology also covers relevant information from other domains, such as individual user profile, biomarker analysis, and dietary reference intakes and recommendations, in order to properly reason with food details in the context of individual users and disease conditions. Having this kind of ontology is a benefit for helping e-health systems to reason with food-composition data in order to better assist users in making the correct, healthy diet choice for particular health condition, age, body mass index, lifestyle, etc. The ontology was created to enable harmonization between the personalized web services that are part of the Qualify project. Having in mind that they are all dealing with food-related data and different ways of describing and classifying data,



(a) Slovenia (b) Denmark (c) The United Kingdom

Fig 6.6: Word clouds of ingredients for which we found analyses in databases of European countries.

the question was how to extract knowledge from them. For this a RESTful web service called "HarmonizedQuisper" that is a part of the QuaLiFY Server Platform (Quisper) was developed. The web service provides a single point of access to the QuaLiFY Server Platform web services. Requests use the purpose-designed Data Query Language (DQL). The DQL is used for defining the information content and the search options in the requests. The language is similar to SQL, but it is not connected with a data model in any specific database; it is based on the Quisper ontology. The reserved words with which the query sentence structure is defined are the concepts presented in the Quisper ontology. The responses depend on which web service is called. The main task where this ontology is used is to find the appropriate web service to which we need to address our input, and using it together with the created taxonomy to translate the input query to the appropriate query for the web service that needs to be requested.

6.4 Conclusion

In this chapter, to see the contribution of the proposed methods to nutrition science, we provided the reasons why these methods were developed together with a discussion of possible real-world application that can benefit of using them. Our third hypothesis is that all of the proposed methods can improve quality of some analyses performed in nutrition science. The hypothesis was evaluated by providing a discussion of a real-world application in state-of-art nutrition science. The discussion contains illustrative examples and future applications that will benefit of using them. Each of the proposed methods is crucial for nutrition science and is developed to solve different task. Further, the results obtained from them can be combined for some more advanced analysis, such as automatic dietary assessment or to plan and implement national food and nutrition programmes and interventions.

Chapter 7

Conclusions and Further Work

Nowadays, numerous research studies involve data that needs to be analyzed. One of the most common ways of performing data analysis is to use statistical analysis to correctly interpret the results. This is especially important working with experimental data in biology, chemistry, physics, social sciences, etc. In such cases, descriptive statistics are used to summarize the information about some parameter that is measured during the experiment, or even more, statistical comparison is often used to find the statistical significance between different methods that are used to measure the same parameter over different problems, or to find if there is a statistical significance between different groups for which the same parameter is measured in different conditions. These experiments are usually performed by obtaining multiple independent measurements on each problem, which can further be analyzed in one of the two possible scenarios: single-problem or multiple-problem scenario. In the case of the single-problem scenario, data sets that contain experimental data are analyzed by providing simple statistics such as averages or medians, or performing a standard omnibus statistical test for comparing averages, medians, or variances. In the case of multiple-problem scenario, the data sets are compared over different problems, so the commonly-used approach from classical inferential statistics uses average or median of the multiple measurements obtained on the same problem as the representative value for that problem, which are further analyzed by applying a standard omnibus statistical test. The question that arises is if the information about averages or medians, or using them as representative values in the multiple-problem scenario, suffices for good interpretation of the results. For example, averaging is sensitive to outliers that can exist during some experiments regarding their uncertainty and they also affect the end result of the statistical analysis. Medians can be used as more robust statistics than averages, but it can happen that both, averages and medians, are sensitive to some statistically insignificant differences, which results from the ranking scheme used by some standard omnibus statistical tests that can be solved if we give the prior information about the magnitude of the difference for which the data range needs to be known. Another problem can be that the data range can change for the same parameter and different problems. Even more, the medians can be the same but the distributions of the data can be different (e.g., two Gaussian distributions with different variances). All of these factors need to be considered with great care when performing a statistical analysis in order to provide more reliable results of the study. To overcome all these factors that influence the end result of statistical analyses, we propose a novel approach, which removes the sensitivity of simple statistics to outliers and statistically insignificant difference provided by ranking schemes of some standard statistical tests. The approach also defines automatically the magnitude of the insignificant difference for different data ranges. The main benefit of using it is that more robust statistics are obtained that are further used as an input to a standard omnibus statistical test. The

data involved in the comparison is going under transformation by using a novel ranking scheme based on comparing distributions, so the transformed data is sensitive to the shape and location of the distribution. Using experimental data, we proved that Deep Statistical Comparison approach gives a more robust result than the commonly-used approach from classical inferential statistics, when the results are affected by the presence of outliers and statistically insignificant differences [25]–[29]. When the results are not affected, both approaches give the same result. Robustness of the results is crucial in nutrition science because it is important for establishing dietary recommendations.

In the second part of the thesis we addressed the problem of information extraction for nutrition science. The results of a study are usually presented in scientific publications. To follow the new knowledge in a particular domain that comes rapidly with new published papers or in any other electronic form, NER methods are used to automatically identify and extract the relevant information in the domain, which can be further used for more advanced analysis. For this reason, in the last 10 years, the NLP community organized shared workshops for information extraction from biomedical literature, which is a domain composed from several subdomains including gene entity recognition, chemical entity recognition, drug entity recognition, disease entity recognition, etc. The workshops provided annotated corpora for the domain and in each one most of the presented methods are corpus-based NERs based on ML algorithms. However, having an annotated corpus in an untapped domain is a time-consuming task, which also requires a good knowledge of the domain, so human experts should be involved in order to produce a good corpus. In such cases, rule-based NERs can be used, in which rules are created using the characteristics of the entities of interest, which also requires a good knowledge of the domain. Our work is focused on the dietary domain, which is related to the biomedical domain and it is still an untapped domain. The focus on the domain comes from two European initiatives, ENPADASI (`http://www.enpadasi.eu/`) that started in mid December 2014, and RICHFIELDS (`https://www.richfields.eu/`) that started in Autumn 2015, in which the main objective is to deliver an open access research infrastructure that will contain data from a wide variety of nutritional studies and food behaviors, respectively. To extract data from the dietary domain, we propose a rule-based NER. The difference with the other rule-based NERs is that we do not use rules associated with the characteristics of the entities of interest but we use a small number of Boolean algebra rules that help us define the phrases that are entities mentions [30]–[32]. Experiments have shown that the presented method provides good results when an annotated corpus is missing and there are no rules created by human experts, so the presented methodology can be easily extended to some other untapped domain. After the extraction of the information of interest, the same concept may be mentioned using phrases with a variety of structures, which is a consequence of how people express themselves. This is a problem especially in the cases when the information about the same concept needs to be combined for a further data analysis.

To overcome this problem, text normalization methods exist in which the main aim is to link the concept to known concept from a domain-specific terminological resource. Text normalization methods are based on string similarity measures that can fail when they are dealing with short segments of text. Because the extracted entities are usually presented as short segments of text, we propose a new method for a string similarity of domain-specific short segments of text. The methodology that is presented uses the probability theory to model the domain using the morphological information that is presented in the text. Comparing the method with some standard string similarity measures that are typically used in text normalization methods, the proposed method provides the best result in the food matching problem [33]–[38]. The method can be easily extended to each domain. An example of system that uses this method is a StandFood, which is a semi-automatic system

for classifying and describing foods according to FoodEx2 [204] and can be used for food data harmonization.

In this thesis, we presented an approach for exploring a given domain. The approach brings together a synergy of statistics and natural language processing. Statistics is focused on how to obtain more robust results that need to be published, and natural language processing is used to extract the relevant scientifically published information in order to follow the new knowledge from the domain.

From statistics, we focused on hypothesis testing. We proposed a novel method for making a statistical comparison of experimental data that is more robust on outliers and small differences that can exist between the data values. The method introduces a new ranking scheme, which uses the whole distribution of the data instead of using only one statistic to describe the data distribution, such as average or median. Using it, we improved the results obtained from statistical comparisons, especially in cases when the comparisons are affected by the presence of outliers or by small ϵ-neighbourhood that exists between the data values. The approach is called *Deep Statistical Comparison (DSC)* and it is generally applicable in any domain such as nutrition science, chemistry, biology, physics, computational intelligence, machine learning, social sciences, etc.

From the perspective of natural language processing, we focused on information extraction and text normalization. We proposed a rule-based named-entity recognition method (NER) that can be used to extract knowledge from an untapped domain. The method does not require an annotated corpus and we extended the idea of rule-based NERs in such a way that the rules are not associated with the characteristics of the entities of interest. The extracted information (entities) may be mentioned in text using phrases with a variety of structures. To allow automatic integration of the information available for the same entity that is represented in different ways text normalization methods are used, which are based on string similarity measures. For this reason, we proposed a novel method for string similarity of domain-specific short segments of text, called *Part-Of-Speech tagging probability weighted method*. The method gives a probability model of the domain using the morphological information presented in the text. Using it, we improved the string similarity that can be achieved using some standard string similarity measures.

The evaluation of the methods is made in the nutrition science domain. Statistics is a crucial part in public health because dietary recommendation should rely on evidence-based principles, considering scientific knowledge, expert consensus, and clinical experience that are proved with statistical analyses. The evaluation of the DSC proposed method shows that it gives more robust results than some standard statistical comparison methods. The results of statistical analyses are then published in scientific publications. To follow the new knowledge about the dietary domain that is massive and is quickly increasing, named-entity recognition methods are welcome to support knowledge identification, extraction and exploration. For this reason, we used drNER, which is a rule-based NER. The additional challenge here is that the dietary domain is an untapped domain, so resources such as annotated corpus, ontologies, etc., are still missing. From its evaluation, we show that promising results can be achieved. Finally, to link the same concepts that are represented using different phrases or to applied text normalization, the POS tagging probability weighted method is applied on food concepts, or the well-known food matching problem. From its evaluation, we show that the method gives more promising results compared to some standard string similarity measure applied on the same problem.

During the work and the evaluation of the proposed methods that are part of this thesis, we have identified several open questions that lead to the following directions for further work.

The first direction is related to the DSC method. Working with food composition

databases (FCDBs) there are a lot of missing values. Nowadays, if we are using a food composition table from some country and there is a missing value for some nutrition parameter for a given food item that belongs to a given food group, manually-created rules exist that defined from which countries the nutrition parameter value should be borrowed. Further, we will use the idea of DSC to compare different countries according to a nutrition parameter value in a given food group. According to the statistical results we can define rules that can be used to missing values imputation and they will be further compared with the manually-created rules. The DSC method works with one-dimensional data, however the results of some experiments are high-dimensional data that also need to be compared using statistical analyses. For this reason, we will further focus on extending the DSC method in order to apply it to high-dimensional data. Also, we will further focus on comparing the DSC approach with the standard approach using artificial data including imputed artificial outliers in order to check that the approach is insensitive to outliers and to quantify the robustness. One direction is also to change and make DSC available from perspective of Bayesian statistics.

The second direction is related to information extraction and text normalization. The dietary domain is an untapped domain. The drNER was the first attempt in which we tried to extract information from it. Further, we will focus on finding a good way to knowledge representation and management of the dietary domain, by exploring the existing ontologies or learning a new one, which is also a goal of the H2020 European project RICHFIELDS. Then, by using the proposed ontology, we will further try to build an annotated corpus that will be used to introduce AL NERS or corpus-based NERs to improve the results for information extraction in the dietary domain. Further, using the extracted information in the dietary domain and the information about drug, diseases, and genes that can be obtained from NERs presented as part of some previous BioNLP workshops, we will focus on extracting the relations that exist between the extracted entities. Finally, the proposed text normalization method will be used for normalization of the extracted entities in order to help the process of automatic integration of the information that exists for the same entity.

The third direction is related to nutrition science. We will further use drNER to produce data sets that will contain the last dietary recommendations according to a specific selected disease.

Appendix A

Two-Sample Kolmogorov-Smirnov Test

The two-sample *Kolmogorov-Smirnov test*, also known as the two-sample KS test, is used to test whether two samples come from the same distribution. It is very similar to the one-sample *Kolmogorov-Smirnov test*, and *Kolmogorov-Smirnov test* for normality [45].

Let us suppose that a first independent and identically distributed (i.i.d) sample $\mathbf{X_1^{(1)}}$, $\mathbf{X_2^{(1)}}, \ldots, \mathbf{X_m^{(1)}}$ of size m has distribution with cumulative distribution function (c.d.f.) $F_{1,m}(x)$, and the second i.i.d sample $\mathbf{X_1^{(2)}}, \mathbf{X_2^{(2)}}, \ldots, \mathbf{X_n^{(2)}}$ of size n has distribution with cumulative distribution function (c.d.f.) $F_{2,n}(x)$. The hypotheses we want to test are

$$\begin{aligned} H_0 &: F_{1,m} = F_{2,n}, \\ H_1 &: F_{1,m} \neq F_{2,n}. \end{aligned} \tag{A.1}$$

In this case, the KS test statistic is

$$\mathbf{D_{m,n}} = \sup_x |F_{1,m}(x) - F_{2,n}(x)|, \tag{A.2}$$

where *sup* is the supremum function.

The null hypothesis is rejected at a significance level α if

$$D_{m,n,\alpha} > c(\alpha)\sqrt{\frac{m+n}{mn}}, \tag{A.3}$$

where the value $c(\alpha)$ is given in Kolmogorov-Smirnov table for each significance level α.

Appendix B

Data Used for Statistical Experiments

The proposed approach is applicable to any domain. Because we lack of raw experimental food compisition data to evaluate the novel proposed statistical approach, we used the results from the Black-Box Benchmarking 2015 (BBOB 2015) [228] competition. BBOB 2015 is a competition that provides single-objective functions for benchmarking, which are then optimized by competing algorithms. From the competition, 15 out of 18 algorithms were used for evaluation because 3 of the algorithms did not provide data organized in the template provided by the BBOB 2015. The algorithms used were: BSif [229], BSifeg [229], BSqi [229], BSrr [229], CMA-CSA [230], CMA-MSR [230], CMA-TPA [230], GP1-CMAES [231], GP5-CMAES [231], RAND-2xDefault [232], RF1-CMAES [231], RF5-CMAES [231], Sif [229], Sifeg [229], and Srr [229]. For our propose and the idea of experiments in nutrition science, we renamed the algorithms as countries, $C_i, i = 1, \ldots, 15$, using an alphabetical order. For each one, the results for 24 different noiseless test functions in 5 dimensionality (2, 3, 5, 10, and 20) were selected. For the experiments, a statistical comparison was performed by comparing the algorithms on 22 different noiseless functions since not all algorithms provided data for the two benchmark functions for some dimensions. For each algorithm, the BBOB 2015 provided data for 15 runs on each problem. The test functions, F, are from 5 groups that represent

- separable functions,
- functions with low or moderate conditioning,
- functions with high conditioning and unimodal,
- multi-modal functions with an adequate global structure, and
- multi-modal functions with a week global structure.

More details about them can be found in [233].

Appendix C

Rules for post-hoc chunkings

On each sentence segment only three rules were applied, one per each post-hoc chunking.

The rule for the first post-hoc chunking is a Boolean *AND* function, which defines when a trigram of successive chunks composed as $(B-NP, B-PP, B-NP)$ needs to be merged into a new noun chunk. In our case, the function consists of six expressions, which are defined for all variations without repetition, $V_{3,2} = 6$. For each variation, the expression is a Boolean *NAND* function defined as

$$f(A, B) = \neg A \vee \neg B = \neg(A \wedge B), \tag{C.1}$$

where A and B are different entities.

The rule for the second post-hoc chunking only checks if a trigram of successive chunks is composed as (*B-NP*, *B-VP*, *B-NP*) and the first noun chunk has a POS tag that is Wh-pronoun such as who, what, which, etc., in order to merge them into a new noun chunk.

The rule for the third post-hoc chunking is a Boolean *AND* function, which defines when a bigram of successive chunks composed as $(B - NP, B - NP)$ needs to be merged into a new noun chunk. In our case, the function consists of nine expressions. Three of them are related when both noun phrases have the same entity level, $C_{3,1} = 3$, and six correspond to the cases when only one of the noun phrases is labeled as entity of interest, so first we need to select which one and then to select which entity will be on that place, $C_{2,1} \cdot C_{3,1} = 6$.

References

[1] H. Vorster, P. Love, and C. Browne, "Development of food-based dietary guidelines for south africa: the process," *South African Journal of Clinical Nutrition*, vol. 14, no. 3, 2001.

[2] J. J. Otten, J. P. Hellwig, L. D. Meyers, *et al.*, *Dietary reference intakes: the essential guide to nutrient requirements*. National Academies Press, 2006.

[3] EFSA, *European food safety authority*, accessed September 18, 2016. [Online]. Available: `https://www.efsa.europa.eu/`.

[4] EURRECA, *European micronutrient recommendations aligned*, accessed September 18, 2016. [Online]. Available: `http://www.ilsi.org/Europe/Pages/EC_EURRECA.aspx`.

[5] E. L. Doets, L. S. de Wit, R. A. Dhonukshe-Rutten, A. E. Cavelaars, M. M. Raats, L. Timotijevic, A. Brzozowska, T. M. Wijnhoven, M. Pavlovic, T. H. Totland, *et al.*, "Current micronutrient recommendations in europe: towards understanding their differences and similarities," *European Journal of Nutrition*, vol. 47, no. 1, pp. 17–40, 2008.

[6] EuroFIR, *European food information resource*, accessed September 18, 2016. [Online]. Available: `http://www.eurofir.org/`.

[7] QuaLiFY, *Information service for personalised nutrition and lifestyle advice*, accessed September 18, 2016. [Online]. Available: `http://quisper.eu/`.

[8] Hyve and TNO, *Food4Me*, accessed September 18, 2016. [Online]. Available: `http://food4me.org`.

[9] H. Greenfield and D. Southgate, *Food composition data: production, management, and use*. Food & Agriculture Org., 2003.

[10] G. Spaargaren, P. Oosterveer, and A. Loeber, *Food practices in transition: changing food consumption, retail and production in the age of reflexive modernity*. Routledge, 2013.

[11] F. B. Hu, "Dietary assessment methods," *Obesity Epidemiology,*, pp. 108–111, 2008.

[12] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[13] S. García and F. Herrera, "An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons," *Journal of Machine Learning Research*, vol. 9, no. Dec, pp. 2677–2694, 2008.

[14] S. García, D. Molina, M. Lozano, and F. Herrera, "A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: a case study on the cec2005 special session on real parameter optimization," *Journal of Heuristics*, vol. 15, no. 6, pp. 617–644, 2009.

[15] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power," *Information Sciences*, vol. 180, no. 10, pp. 2044–2064, 2010.

[16] J. Derrac, S. García, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 3–18, 2011.

[17] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[18] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS biology*, vol. 2, no. 11, e309, 2004.

[19] A. B. do Nascimento, G. M. R. Fiates, A. dos Anjos, and E. Teixeira, "Analysis of ingredient lists of commercially available gluten-free and gluten-containing food products using the text mining technique," *International journal of food sciences and nutrition*, vol. 64, no. 2, pp. 217–222, 2013.

[20] J. Xia, X. Zhang, D. Yuan, L. Chen, J. Webster, and A. C. Fang, "Gene prioritization of resistant rice gene against xanthomas oryzae pv. oryzae by using text mining technologies," *BioMed research international*, vol. 2013, 2013.

[21] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[22] N. Alnazzawi, P. Thompson, R. Batista-Navarro, and S. Ananiadou, "Using text mining techniques to extract phenotypic information from the phenochf corpus," *BMC medical informatics and decision making*, vol. 15, no. 2, p. 1, 2015.

[23] R. Leaman, C.-H. Wei, C. Zou, and Z. Lu, "Mining patents with tmchem, gnormplus and an ensemble of open systems," in *Proce. The fifth BioCreative challenge evaluation workshop*, 2015, pp. 140–146.

[24] T. Arnold and L. Tilton, "Natural language processing," in *Humanities Data in R*, Springer, 2015, pp. 131–155.

[25] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Disadvantages of statistical comparison of stochastic optimization algorithms," *Proceedings of the Bioinspired Optimizaiton Methods and their Applications, BIOMA*, pp. 105–118, 2016.

[26] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Deep statistical comparison applied on quality indicators to compare multi-objective stochastic optimization algorithms," *Proceedings of the Third International Conference on Machine Learning, Optimization and Big Data, MOD 2017*, 2017.

[27] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "The behavior of deep statistical comparison approach for different criteria of comparing distirbutions," in *Proceedings of the 9th International Conference on Computational Intellifence, (IJCCI 2017)*, vol. 1, 2017, pp. 73–82.

[28] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Comparing multi-objective optimization algorithms using an ensemble of quality indicators with deep statistical comparison approach," *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI 2017)*, pp. 2801–2809, 2017.

[29] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "A novel approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics," *Information Sciences*, vol. 417, pp. 186–215, 2017.

[30] T. Eftimov, B. Koroušić Seljak, and P. Korošec, "Grammar and dictionary based named-entity linking for knowledge extraction of evidence-based dietary recommendations," in *Proceedings of the 8th international Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, (IC3K 2016)*, vol. 1:KDIR, 2016, pp. 150–157.

[31] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Knowledge extraction of evidence - based dietary recommendations," in *Proceedings of the Jožef Stefan International Postgraduate School Students Conference 2016*, Jožef Stefan International Postgraduate School, vol. 1, 2016, pp. 149–153.

[32] T. Eftimov, B. Koroušić Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," *PloS One*, vol. 12, no. 6, e0179488, 2017.

[33] T. Eftimov and B. Koroušić Seljak, "POS tagging-probability weighted method for matching the internet recipe ingredients with food composition data," in *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, IEEE, vol. 1, 2015, pp. 330–336.

[34] T. Eftimov, S. Mezgec, P. Korošec, and B. Koroušić Seljak, "Mixed deep learning and natural language processing approach for food image detection, recognition and analysis to estimate nutritional values," in *Book of abstracts - Advancing Behaviour Change Science*, ISBNPA, 2017.

[35] T. Eftimov, G. Ispirova, P. Korošec, and B. Koroušić Seljak, "A semi-automatic system for classifying and describing foods according to FoodEx2," in *3rd IMEKO FOODS, Metrology promoting Standardization and Harmonization in Food and Nutrition*, 2017, pp. 56–59.

[36] T. Eftimov, P. Korošec, B. Koroušić Seljak, M. Roe, A. Turrini, and P. Finglas, "Matching foods from EuroFIR databases to FoodEx by using a semi-automatic system for classifying and describing foods," in *Book of abstracts: 12th International Food Data Conference*, 2017.

[37] G. Ispirova, T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Mapping food composition data from various data sources to a domain-specific ontology," in *Proceedings of the 9th International Conference on Knowledge Engineering and Ontology Development, (KEOD 2017)*, vol. 2, 2017, pp. 203–210.

[38] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Standfood: standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2," *Nutrients*, vol. 9, no. 6, p. 542, 2017.

[39] T. Eftimov, P. Korošec, D. Potočnik, N. Ogrinc, D. Heath, and B. Koroušić Seljak, "How to perform properly statistical analysis on food data? an e-learning tool: advanced statistics in natural sciences and technologies," *Book chapter - Science within Food: Up-to-date Advances on Research and Educational Ideas*, 2017.

[40] R. G. Steel, J. H. Torrie, *et al.*, "Principles and procedures of statistics.," *Principles and procedures of statistics.*, 1960.

[41] G. S. Wasserman, *Probability and statistics for modern engineering*, 1991.

[42] B. J. Winer, D. R. Brown, and K. M. Michels, *Statistical principles in experimental design*. McGraw-Hill New York, 1971, vol. 2.

[43] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing statistical hypotheses*. Wiley New York et al, 1986, vol. 150.

[44] E. Erdfelder, F. Faul, and A. Buchner, "Gpower: a general power analysis program," *Behavior research methods, instruments, & computers*, vol. 28, no. 1, pp. 1–11, 1996.

[45] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[46] A. Pettitt, "Testing the normality of several independent samples using the anderson-darling statistic," *Applied Statistics*, pp. 156–161, 1977.

[47] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[48] E. S. Pearson, R. B. D "'AGOSTINO, and K. O. Bowman, "Tests for departure from normality: comparison of powers," *Biometrika*, vol. 64, no. 2, pp. 231–246, 1977.

[49] N. M. Razali, Y. B. Wah, *et al.*, "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests," *Journal of statistical modeling and analytics*, vol. 2, no. 1, pp. 21–33, 2011.

[50] B. B. Schultz, "Levene's test for relative variation," *Systematic Zoology*, vol. 34, no. 4, pp. 449–456, 1985.

[51] N. Cressie and H. Whitford, "How to use the two sample t-test," *Biometrical Journal*, vol. 28, no. 2, pp. 131–148, 1986.

[52] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[53] H. Hsu and P. A. Lachenbruch, "Paired t test," *Wiley Encyclopedia of Clinical Trials*, 2008.

[54] F. Lam and M. Longnecker, "A modified wilcoxon rank sum test for paired data," *Biometrika*, vol. 70, no. 2, pp. 510–513, 1983.

[55] M. R. Stoline, "The status of multiple comparisons: simultaneous estimation of all pairwise comparisons in one-way anova designs," *The American Statistician*, vol. 35, no. 3, pp. 134–141, 1981.

[56] N. Breslow, "A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship," *Biometrika*, vol. 57, no. 3, pp. 579–594, 1970.

[57] E. R. Girden, *ANOVA: Repeated measures*, 84. Sage, 1992.

[58] P. Nemenyi, "Distribution-free multiple comparisons," in *PhD Thesis*, Princeton University, 1963.

[59] O. J. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.

[60] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.

[61] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.

[62] B. Bergmann and G. Hommel, "Improvements of general multiple test procedures for redundant systems of hypotheses," in *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, Springer, 1988, pp. 100–115.

[63] J. Piskorski and R. Yangarber, "Information extraction: past, present and future," in *Multi-source, multilingual information extraction and summarization*, Springer, 2013, pp. 23–49.

[64] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[65] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT Press, 1999, vol. 999.

[66] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.

[67] J. Plisson, N. Lavrac, and D. Mladenic, "A rule based approach to word lemmatization," *Proceedings of IS-2004*, pp. 83–86, 2004.

[68] J. B. Lovins, *Development of a stemming algorithm*. MIT Information Processing Group, Electronic Systems Laboratory Cambridge, 1968.

[69] D. A. Hull, "Stemming algorithms: a case study for detailed evaluation," *JASIS*, vol. 47, no. 1, pp. 70–84, 1996.

[70] A.Voutilainen, "Part-of-speech tagging," *The Oxford handbook of computational linguistics*, pp. 219–232, 2003.

[71] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of the international conference on new methods in language processing*, Citeseer, vol. 12, 1994, pp. 44–49.

[72] Y. Tian and D. Lo, "A comparative study on the effectiveness of part-of-speech tagging techniques on bug reports," in *Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on*, IEEE, 2015, pp. 570–574.

[73] B. Santorini, "Part-of-speech tagging guidelines for the penn treebank project (3rd revision)," 1990.

[74] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: the penn treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[75] A. Taylor, M. Marcus, and B. Santorini, "The penn treebank: an overview," in *Treebanks*, Springer, 2003, pp. 5–22.

[76] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, 2006.

[77] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, 1995, pp. 1137–1145.

[78] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[79] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[80] S. R. Eddy, "Hidden markov models," *Current opinion in structural biology*, vol. 6, no. 3, pp. 361–365, 1996.

[81] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the eighteenth international conference on machine learning, ICML*, vol. 1, 2001, pp. 282–289.

[82] A. McCallum, D. Freitag, and F. C. Pereira, "Maximum entropy markov models for information extraction and segmentation.," in *Icml*, vol. 17, 2000, pp. 591–598.

[83] T. G. Dietterich, "Ensemble learning," *The handbook of brain theory and neural networks*, vol. 2, pp. 110–125, 2002.

[84] R. A. Miller, F. M. Gieszczykiewicz, J. K. Vries, and G. F. Cooper, "Chartline: providing bibliographic references relevant to patient charts using the umls metathesaurus knowledge sources.," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, American Medical Informatics Association, 1992, p. 86.

[85] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program.," in *Proceedings of the AMIA Symposium*, American Medical Informatics Association, 2001, p. 17.

[86] X. Zhou, X. Zhang, and X. Hu, "Maxmatcher: biological concept extraction using approximate dictionary lookup," in *Pacific Rim International Conference on Artificial Intelligence*, Springer, 2006, pp. 1145–1149.

[87] D. Farmakiotou, V. Karkaletsis, J. Koutsias, G. Sigletos, C. D. Spyropoulos, and P. Stamatopoulos, "Rule-based named entity recognition for greek financial texts," in *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, Citeseer, 2000, pp. 75–78.

[88] G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, "Using machine learning to maintain rule-based named-entity recognition and classification systems," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2001, pp. 426–433.

[89] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, "Prominer: rule-based protein and gene entity recognition," *BMC bioinformatics*, vol. 6, no. 1, S14, 2005.

[90] T. Rocktäschel, M. Weidlich, and U. Leser, "Chemspot: a hybrid system for chemical named entity recognition," *Bioinformatics*, vol. 28, no. 12, pp. 1633–1640, 2012.

[91] T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter, "Edgar: extraction of drugs, genes and relations from the biomedical literature," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, NIH Public Access, 2000, p. 517.

[92] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, no. 55-66, p. 11, 2010.

[93] Y. Chen, S. Mani, and H. Xu, "Applying active learning to assertion classification of concepts in clinical text," *Journal of biomedical informatics*, vol. 45, no. 2, pp. 265–272, 2012.

[94] Y. Chen, R. J. Carroll, E. R. M. Hinz, A. Shah, A. E. Eyler, J. C. Denny, and H. Xu, "Applying active learning to high-throughput phenotyping algorithms for electronic health records data," *Journal of the American Medical Informatics Association*, vol. 20, no. e2, e253–e259, 2013.

[95] Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, and H. Xu, "A study of active learning methods for named entity recognition in clinical text," *Journal of biomedical informatics*, vol. 58, pp. 11–18, 2015.

[96] Y. Chen, T. A. Lask, Q. Mei, Q. Chen, S. Moon, J. Wang, K. Nguyen, T. Dawodu, T. Cohen, J. C. Denny, *et al.*, "An active learning-enabled annotation system for clinical named entity recognition," *BMC medical informatics and decision making*, vol. 17, no. 2, p. 82, 2017.

[97] V. C. Tran, N. T. Nguyen, H. Fujita, D. T. Hoang, and D. Hwang, "A combination of active learning and self-learning for named entity recognition on twitter using conditional random fields," *Knowledge-Based Systems*, vol. 132, pp. 179–187, 2017.

[98] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, ACM, 2008, pp. 160–167.

[99] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[100] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," *arXiv preprint arXiv:1511.08308*, 2015.

[101] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[102] C. N. d. Santos and V. Guimaraes, "Boosting named entity recognition with neural character embeddings," *arXiv preprint arXiv:1505.05008*, 2015.

[103] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv preprint arXiv:1603.01360*, 2016.

[104] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.

[105] M. M. Lopez and J. Kalita, "Deep learning applied to nlp," *arXiv preprint arXiv:1703.03091*, 2017.

[106] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," *arXiv preprint arXiv:1707.05928*, 2017.

[107] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, 2013.

[108] D. Metzler, S. Dumais, and C. Meek, "Similarity measures for short segments of text," in *European Conference on Information Retrieval*, Springer, 2007, pp. 16–27.

[109] P. A. Hall and G. R. Dowling, "Approximate string matching," *ACM computing surveys (CSUR)*, vol. 12, no. 4, pp. 381–402, 1980.

[110] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[111] W. E. Winkler, "String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage.," 1990.

[112] A. Barrón-Cedeno, P. Rosso, E. Agirre, and G. Labaka, "Plagiarism detection across distant language pairs," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 37–45.

[113] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, "A closer look at skip-gram modelling," in *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, sn, 2006, pp. 1–4.

[114] E. F. Krause, *Taxicab geometry: An adventure in non-Euclidean geometry*. Courier Corporation, 2012.

[115] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[116] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[117] J. Gill, "The insignificance of null hypothesis significance testing," *Political Research Quarterly*, vol. 52, no. 3, pp. 647–674, 1999.

[118] T. R. Levine, R. Weber, C. Hullett, H. S. Park, and L. L. M. Lindsey, "A critical assessment of null hypothesis significance testing in quantitative communication research," *Human Communication Research*, vol. 34, no. 2, pp. 171–187, 2008.

[119] N. R. Schooler, "The statistical comparison of clinical trials.," *The Journal of clinical psychiatry*, vol. 62, pp. 35–7, 2001.

[120] N. T. Bailey, *Statistical methods in biology*. Cambridge university press, 1995.

[121] P. C. Meier and R. E. Zünd, *Statistical methods in analytical chemistry*. John Wiley & Sons, 2005, vol. 211.

[122] J. Gibbons and S. Chakraborti, *International encyclopedia of statistical science*. 2011.

[123] J. H. McDonald, *Handbook of biological statistics*. Sparky House Publishing Baltimore, MD, 2009, vol. 2.

[124] Ö. Senger and A. K. Çelik, "A monte carlo simulation study for kolmogorov-smirnov two-sample test under the precondition of heterogeneity: upon the changes on the probabilities of statistical power and type i error rates with respect to skewness measure," *Journal of Statistical and Econometric Methods*, vol. 2, no. 4, pp. 1–16, 2013.

[125] M. J. Laan van der, S. Dudoit, and K. S. Pollard, "Multiple testing. part ii. step-down procedures for control of the family-wise error rate," *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, pp. 1–33, 2004.

[126] C. Z. Mooney, *Monte carlo simulation*. Sage Publications, 1997, vol. 116.

[127] G. Keppel and T. Wickens, "Simultaneous comparisons and the control of type i errors," *Design and analysis: A researcher's handbook. 4th ed. Upper Saddle River (NJ): Pearson Prentice Hall. p*, pp. 111–130, 2004.

[128] M. Rosenblatt, "A central limit theorem and a strong mixing condition," *Proceedings of the National Academy of Sciences*, vol. 42, no. 1, pp. 43–47, 1956.

[129] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003.

[130] J. H. Zar *et al.*, *Biostatistical analysis*. Pearson Education India, 1999.

[131] R Core Team, *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: `https://www.R-project.org/`.

[132] J. L. Gastwirth, Y. R. Gel, W. L. Wallace Hui, V. Lyubchich, W. Miao, and K. Noguchi, *Lawstat: tools for biostatistics, public policy, and law*, R package version 3.0, 2015. [Online]. Available: `https://CRAN.R-project.org/package=lawstat`.

[133] B. Calvo and G. Santafe, "Scmamp: statistical comparison of multiple algorithms in multiple problems," *The R Journal*, vol. Accepted for publication, 2015.

[134] G. Hommel, "A stagewise rejective multiple test procedure based on a modified bonferroni test," *Biometrika*, vol. 75, no. 2, pp. 383–386, 1988.

[135] S. Engmann and D. Cousineau, "Comparing distributions: the two-sample anderson-darling test as an alternative to the kolmogorov-smirnoff test," *Journal of Applied Quantitative Methods*, vol. 6, no. 3, pp. 1–17, 2011.

[136] D. Fiser, S. Pollak, and S. Vintar, "Learning to mine definitions from slovene structured and unstructured knowledge-rich resources.," in *LREC*, 2010.

[137] S. Pollak, A. Vavpetic, J. Kranjc, N. Lavrac, and S. Vintar, "Nlp workflow for online definition extraction from english and slovene text corpora.," in *KONVENS*, 2012, pp. 53–60.

[138] R. Stanković, C. Krstev, I. Obradović, B. Lazić, and A. Trtovac, "Rule-based automatic multi-word term extraction and lemmatization," in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC*, 2016, pp. 507–514.

[139] N. A. L. United States Department of Agriculture, *Food and nutrition information center*, 2016. [Online]. Available: `https://fnic.nal.usda.gov/`.

[140] C. Nédellec, R. Bossy, J.-D. Kim, J.-J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, "Overview of bionlp shared task 2013," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, 2013, pp. 1–7.

[141] E. Chaix, B. Dubreucq, A. Fatihi, D. Valsamou, R. Bossy, M. Ba, L. Deléger, P. Zweigenbaum&, P. Bessieres, L. Lepiniec, *et al.*, "Overview of the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016," *ACL 2016*, p. 1, 2016.

[142] Y. Luo, Ö. Uzuner, and P. Szolovits, "Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations," *Briefings in bioinformatics*, vol. 18, no. 1, pp. 160–178, 2017.

[143] J.-D. Kim, Y. Wang, N. Colic, S. H. Baek, Y. H. Kim, and M. Song, "Refactoring the genia event extraction shared task toward a general framework for ie-driven kb development," *ACL 2016*, p. 23, 2016.

[144] C. Li, Z. Rao, and X. Zhang, "Litway, discriminative extraction for different bio-events," *ACL 2016*, p. 32, 2016.

[145] H. V. Cook, E. Pafilis, and L. J. Jensen, "A dictionary-and rule-based system for identification of bacteria and habitats in text," *ACL 2016*, p. 50, 2016.

[146] J. Lever and S. J. Jones, "Verse: event and relation extraction in the bionlp 2016 shared task," *ACL 2016*, p. 42, 2016.

[147] M. Tiftikci, H. Sahin, B. Büyüköz, A. Yayıkçı, and A. Ozgür, "Ontology-based categorization of bacteria and habitat entities using information retrieval techniques," *ACL 2016*, p. 56, 2016.

[148] F. Mehryary, J. Björne, S. Pyysalo, T. Salakoski, and F. Ginter, "Deep learning with minimal training data: turkunlp entry in the bionlp shared task 2016," *ACL 2016*, p. 73, 2016.

[149] N. C. Panyam, G. Khirbat, K. Verspoor, T. Cohn, and K. Ramamohanarao, "Seedev binary event extraction using svms and a rich feature set," *ACL 2016*, p. 82, 2016.

[150] L. Smith, L. K. Tanabe, R. J. nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, *et al.*, "Overview of biocreative ii gene mention recognition," *Genome biology*, vol. 9, no. 2, p. 1, 2008.

[151] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *Journal of Machine Learning Research*, vol. 6, no. Nov, pp. 1817–1853, 2005.

[152] C.-J. Kuo, Y.-M. Chang, H.-S. Huang, K.-T. Lin, B.-H. Yang, Y.-S. Lin, C.-N. Hsu, and I.-F. Chung, "Rich feature set, unification of bidirectional parsing and dictionary filtering for high f-score gene mention tagging," in *Proceedings of the second BioCreative challenge evaluation workshop*, Centro Nacional de Investigaciones Oncologicas (CNIO) Madrid, Spain, vol. 23, 2007, pp. 105–107.

[153] H.-S. Huang, Y.-S. Lin, K.-T. Lin, C.-J. Kuo, Y.-M. Chang, B.-H. Yang, I.-F. Chung, and C.-N. Hsu, "High-recall gene mention recognition by unification of multiple backward parsing models," in *Proceedings of the second BioCreative challenge evaluation workshop*, Centro Nacional de Investigaciones Oncologicas (CNIO) Madrid, Spain, vol. 23, 2007, pp. 109–111.

[154] R. Klinger, C. M. Friedrich, J. Fluck, and M. Hofmann-Apitius, "Named entity recognition with combinations of conditional random fields," in *Proc. of the Second BioCreative Challenge Evaluation Workshop*, 2007, pp. 89–91.

[155] R. Leaman and G. Gonzalez, "Banner: an executable survey of advances in biomedical named entity recognition.," in *Pacific symposium on biocomputing*, vol. 13, 2008, pp. 652–663.

[156] C. N. Arighi, Z. Lu, M. Krallinger, K. B. Cohen, W. J. Wilbur, A. Valencia, L. Hirschman, and C. H. Wu, "Overview of the biocreative iii workshop," *BMC bioinformatics*, vol. 12, no. 8, p. 1, 2011.

[157] C. N. Arighi, C. H. Wu, K. B. Cohen, L. Hirschman, M. Krallinger, A. Valencia, Z. Lu, J. W. Wilbur, and T. C. Wiegers, "Biocreative-iv virtual issue," *Database*, vol. 2014, bau039, 2014.

[158] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, and A. Valencia, "Chemdner: the drugs and chemical names extraction challenge," *Journal of cheminformatics*, vol. 7, no. 1, p. 1, 2015.

[159] M. Krallinger, O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, D. M. Lowe, *et al.*, "The chemdner corpus of chemicals and drugs and its annotation principles," *Journal of cheminformatics*, vol. 7, no. 1, p. 1, 2015.

[160] T. Nunes, D. Campos, S. Matos, and J. L. Oliveira, "Becas: biomedical concept recognition services and visualization," *Bioinformatics*, btt317, 2013.

[161] M. Bada, L. E. Hunter, M. Eckert, and M. Palmer, "An overview of the craft concept annotation guidelines," in *Proceedings of the Fourth Linguistic Annotation Workshop*, Association for Computational Linguistics, 2010, pp. 207–211.

[162] T. Ohta, S. Pyysalo, J. Tsujii, and S. Ananiadou, "Open-domain anatomical entity mention detection," in *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, Association for Computational Linguistics, 2012, pp. 27–36.

[163] R. I. Doğan, R. Leaman, and Z. Lu, "Ncbi disease corpus: a resource for disease name recognition and concept normalization," *Journal of biomedical informatics*, vol. 47, pp. 1–10, 2014.

[164] D. Campos, S. Matos, and J. L. Oliveira, "Chemical name recognition with harmonized feature-rich conditional random fields," in *BioCreative Challenge Evaluation Workshop*, vol. 2, 2013, p. 82.

[165] S. Kim, R. I. Doğan, A. Chatr-Aryamontri, C. S. Chang, R. Oughtred, J. Rust, R. Batista-Navarro, J. Carter, S. Ananiadou, S. Matos, *et al.*, "Biocreative v bioc track overview: collaborative biocurator assistant task for biogrid," *Database*, vol. 2016, baw121, 2016.

[166] Q. Wang, S. S. Abdul, L. Almeida, S. Ananiadou, Y. I. Balderas-Mart´nez, R. Batista-Navarro, D. Campos, L. Chilton, H.-J. Chou, G. Contreras, *et al.*, "Overview of the interactive task in biocreative v," *Database*, vol. 2016, baw119, 2016.

[167] C.-H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wiegers, and Z. Lu, "Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task," *Database*, vol. 2016, baw032, 2016.

[168] D. C. Comeau, R. I. Doğan, P. Ciccarese, K. B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, F. Rinaldi, M. Torii, *et al.*, "Bioc: a minimalist approach to interoperability for biomedical text processing," *Database*, vol. 2013, bat064, 2013.

[169] Q. Zhu, X. Li, A. Conesa, and C. Pereira, "Gram-cnn: a deep learning approach with local context for named entity recognition in biomedical text," *Bioinformatics*, 2017.

[170] Y. Wang, S. Liu, M. Rastegar-Mojarad, L. Wang, F. Shen, F. Liu, and H. Liu, "Dependency and amr embeddings for drug-drug interaction extraction from biomedical literature," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, 2017, pp. 36–43.

[171] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, "A neural network multi-task learning approach to biomedical named entity recognition," *BMC bioinformatics*, vol. 18, no. 1, p. 368, 2017.

[172] Y.-X. Zhou, Y.-N. Lu, Z.-L. Pei, Y. Lu, and X.-L. Ma, "Deep learning based disease mention recognition," in *Computer Science And Artificial Intelligence-Proceedings Of The International Conference On Computer Science And Artificial Intelligence (Csai2016)*, World Scientific, 2017, p. 419.

[173] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki, "Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level," *PLoS computational biology*, vol. 10, no. 1, e1003432, 2014.

[174] K. Hornik, *Opennlp: apache opennlp tools interface*, R package version 0.2-5, 2015. [Online]. Available: `https://CRAN.R-project.org/package=openNLP`.

[175] R. J. Nelson, "Karnaugh m.. the map method for synthesis of combinational logic circuits. transactions of the american institute of electrical engineers, vol. 72 part i (1953), pp. 593–598.," *The Journal of Symbolic Logic*, vol. 20, no. 02, pp. 197–197, 1955.

[176] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, and D. Mladenic, "Triplet extraction from sentences," in *Proceedings of the 10th International Multiconference" Information Society-IS*, 2007, pp. 8–12.

[177] G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf, "The units ontology: a tool for integrating units of measurement in science," *Database*, vol. 2012, bas033, 2012.

[178] M. Courtot, F. Gibson, A. L. Lister, J. Malone, D. Schober, R. R. Brinkman, and A. Ruttenberg, "Mireot: the minimum information to reference an external ontology term," *Applied Ontology*, vol. 6, no. 1, pp. 23–33, 2011.

[179] U. Visser, S. Abeyruwan, U. Vempati, R. P. Smith, V. Lemmon, and S. C. Schürer, "Bioassay ontology (bao): a semantic description of bioassays and high-throughput screening results," *BMC bioinformatics*, vol. 12, no. 1, p. 257, 2011.

[180] A. Wilson and J. Thomas, "Semantic annotation," *Corpus Annotation. Longman, London*, 1997.

[181] T. McEnery and A. Wilson, *Corpus linguistics: An introduction*. Edinburgh University Press, 2001.

[182] P. Rayson, D. Archer, S. Piao, and A. McEnery, "The ucrel semantic analysis system.," 2004.

[183] U. N. L. of Medicine, *Pubmed central (pmc) apis*, 2016. [Online]. Available: `https://www.ncbi.nlm.nih.gov/home/develop/api.shtml`.

[184] D. M. Lowe and R. A. Sayle, "Leadmine: a grammar and dictionary driven approach to entity recognition," *Journal of cheminformatics*, vol. 7, no. 1, p. 1, 2015.

[185] M. Barros and F. M. Couto, "Knowledge representation and management: a linked data perspective," *IMIA Yearbook*, pp. 178–183, 2016.

[186] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[187] C. Friedman, L. Shagina, S. A. Socratous, and X. Zeng, "A web-based version of medlee: a medical language extraction and encoding system," in *Proceedings of the AMIA Annual Fall Symposium*, American Medical Informatics Association, 1996, p. 938.

[188] S. Ramanan, S. Broido, and P. S. Nathan, "Performance of a multi-class biomedical tagger on clinical records.," in *CLEF (Working Notes)*, 2013.

[189] N. Collier, A. Oellrich, and T. Groza, "Concept selection for phenotypes and diseases using learn to rank," *Journal of biomedical semantics*, vol. 6, no. 1, p. 24, 2015.

[190] N. Alnazzawi, P. Thompson, and S. Ananiadou, "Mapping phenotypic information in heterogeneous textual sources to a domain-specific terminological resource," *PloS one*, vol. 11, no. 9, e0162287, 2016.

[191] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, *et al.*, "Overview of biocreative ii gene normalization," *Genome biology*, vol. 9, no. 2, S3, 2008.

[192] Z. Lu, H.-Y. Kao, C.-H. Wei, M. Huang, J. Liu, C.-J. Kuo, C.-N. Hsu, R. T.-H. Tsai, H.-J. Dai, N. Okazaki, *et al.*, "The gene normalization task in biocreative iii," *BMC bioinformatics*, vol. 12, no. 8, S2, 2011.

[193] Y. Tsuruoka, J. McNaught, J. c. Tsujii, and S. Ananiadou, "Learning string similarity measures for gene/protein name dictionary look-up using logistic regression," *Bioinformatics*, vol. 23, no. 20, pp. 2768–2774, 2007.

[194] R. Leaman, R. Islamaj Doğan, and Z. Lu, "Dnorm: disease name normalization with pairwise learning to rank," *Bioinformatics*, vol. 29, no. 22, pp. 2909–2917, 2013.

[195] X. Su and J. A. Gulla, "Semantic enrichment for ontology mapping," Springer.

[196] M. N. K. Boulos, A. Yassine, S. Shirmohammadi, C. S. Namahoot, and M. Brückner, "Towards an "internet of food": food ontologies for the internet of things," *Future Internet*, vol. 7, no. 4, pp. 372–392, 2015.

[197] D. Çelik, "Foodwiki: ontology-driven mobile safe food consumption system," *The Scientific World Journal*, vol. 2015, 2015.

[198] C. Caracciolo, A. Stellato, S. Rajbahndari, A. Morshed, G. Johannsen, Y. Jaques, and J. Keizer, "Thesaurus maintenance, alignment and publication as linked data: the agrovoc use case," *International Journal of Metadata, Semantics and Ontologies*, vol. 7, no. 1, pp. 65–75, 2012.

[199] *Open food facts*, `http://world.openfoodfacts.org/who-we-are`, Accessed: 2016–02-14.

[200] M. Kolchin and D. Zamula, "Food product ontology: initial implementation of a vocabulary for describing food products," in *Proceeding of the 14th Conference of Open Innovations Association FRUCT, Helsinki, Finland*, 2013, pp. 11–15.

[201] C. Snae and M. Brückner, "Foods: a food-oriented ontology-driven system," in *Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on*, IEEE, 2008, pp. 168–176.

[202] R. Real and J.M.Vargas, "The probabilistic basis of jaccard's index of similarity," *Systematic biology*, pp. 380–385, 1996.

[203] B. Cestnik *et al.*, "Estimating probabilities: a crucial task in machine learning.," in *ECAI*, vol. 90, 1990, pp. 147–149.

[204] E. F. S. Authority, *The food classification and description system foodex2 (revision 2)*, `https://www.efsa.europa.eu/`, (accessed February 17, 2017).

[205] G. Pounis, A. Bonanni, E. Ruggiero, A. Di Castelnuovo, S. Costanzo, M. Persichillo, M. Bonaccio, C. Cerletti, G. Riccardi, M. Donati, *et al.*, "Food group consumption in an italian population using the updated food classification system foodex2: results from the italian nutrition & health survey (inhes) study," *Nutrition, Metabolism and Cardiovascular Diseases*, 2017.

[206] M. P. Van der Loo, "The stringdist package for approximate string matching," *The R*, p. 2, 2014.

[207] L. Mullen, *Tokenizers: a consistent interface to tokenize natural language text*, R package version 0.1.4, 2016. [Online]. Available: `https://CRAN.R-project.org/package=tokenizers`.

[208] S. Birot, C. B. Madsen, A. G. Kruizinga, T. Christensen, A. Crépet, and P. B. Brockhoff, "A procedure for grouping food consumption data for use in food allergen risk assessment," *Journal of Food Composition and Analysis*, vol. 59, pp. 111–123, 2017.

[209] M. Gurinović, J. Milešević, A. Kadvan, M. Djekić-Ivanković, J. Debeljak-Martačić, M. Takić, M. Nikolić, S. Ranković, P. Finglas, and M. Glibetić, "Establishment and advances in the online serbian food and recipe data base harmonized with eurofir™ standards," *Food chemistry*, vol. 193, pp. 30–38, 2016.

[210] C. Zhang and Y. Ma, *Ensemble machine learning*. Springer, 2012, vol. 1.

[211] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of machine learning research*, vol. 2, no. Nov, pp. 45–66, 2001.

[212] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[213] Y. Freund, R. E. Schapire, *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96, 1996, pp. 148–156.

[214] T. P. Jurka, L. Collingwood, A. E. Boydstun, E. Grossman, and W. van Atteveldt, "Rtexttools: a supervised learning package for text classification," *The R journal*, vol. 5, no. 1, pp. 6–12, 2013.

[215] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, IEEE, 1999, pp. 41–48.

[216] N. C. Oza, "Online bagging and boosting," in *Systems, man and cybernetics, 2005 IEEE international conference on*, IEEE, vol. 3, 2005, pp. 2340–2345.

[217] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[218] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[219] L. Collingwood and J. Wilkerson, "Tradeoffs in accuracy and efficiency in supervised learning methods," *Journal of Information Technology & Politics*, vol. 9, no. 3, pp. 298–318, 2012.

[220] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proceedings of the tenth international conference on Information and knowledge management*, ACM, 2001, pp. 403–410.

[221] W. J. Fu, A. J. Stromberg, K. Viele, R. J. Carroll, and G. Wu, "Statistics and bioinformatics in nutritional sciences: analysis of complex data in the era of systems biology," *The Journal of nutritional biochemistry*, vol. 21, no. 7, pp. 561–572, 2010.

[222] W. W. Daniel and C. L. Cross, "Biostatistics: a foundation for analysis in the health sciences," 1995.

[223] H. O. Lancaster and E. Seneta, *Chi-square distribution*. Wiley Online Library, 1969.

[224] R. Team, "Rstudio: integrated development for r," *RStudio, Inc., Boston, MA URL http://www. rstudio. com*, 2015.

[225] W. G. Cochran, "Approximate significance levels of the behrens-fisher test," *Biometrics*, vol. 20, no. 1, pp. 191–195, 1964.

[226] S. B. Green and N. J. Salkind, *Using SPSS for Windows and Macintosh: Analyzing and understanding data*. Prentice Hall Press, 2010.

[227] T. Eftimov and B. Koroušić Seljak, "QOL - Quisper Ontology Learning using personalized dietary services," in *JSI technical report, COBISS.SI-ID 29060391*, 2015.

[228] B. Black Box Optimization Competition, *Black-box benchmarking 2015*, `http://coco.gforge.inria.fr/doku.php?id=bbob-2015`, Accessed: 2016-02-01.

[229] P. Pošík and P. Baudiš, "Dimension selection in axis-parallel brent-step method for black-box optimization of separable continuous functions," in *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference*, ACM, 2015, pp. 1151–1158.

[230] A. Atamna, "Benchmarking ipop-cma-es-tpa and ipop-cma-es-msr on the bbob noiseless testbed," in *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference*, ACM, 2015, pp. 1135–1142.

[231] L. Bajer, Z. Pitra, and M. Holeňa, "Benchmarking gaussian processes and random forests surrogate models on the bbob noiseless testbed," in *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference*, ACM, 2015, pp. 1143–1150.

[232] D. Brockhoff, B. Bischl, and T. Wagner, "The impact of initial designs on the performance of matsumoto on the noiseless bbob-2015 testbed: a preliminary study," in *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference*, ACM, 2015, pp. 1159–1166.

[233] N. Hansen, A. Auger, S. Finck, and R. Ros, "Real-parameter black-box optimization benchmarking 2010: experimental setup," 2010.

Bibliography

Publications Related to the Thesis

Journal Articles

T. Eftimov, P. Korošec, and B. Koroušić Seljak, "A novel approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics," *Information Sciences*, vol. 417, pp. 186–215, 2017.

T. Eftimov, B. Koroušić Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," *PloS One*, vol. 12, no. 6, e0179488, 2017.

T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Standfood: standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2," *Nutrients*, vol. 9, no. 6, p. 542, 2017.

Book Chapters

T. Eftimov, P. Korošec, D. Potočnik, N. Ogrinc, D. Heath, and B. Koroušić Seljak, "How to perform properly statistical analysis on food data? an e-learning tool: advanced statistics in natural sciences and technologies," *Book chapter - Science within Food: Up-to-date Advances on Research and Educational Ideas*, 2017.

Conference Papers

T. Eftimov and B. Koroušić Seljak, "POS tagging-probability weighted method for matching the internet recipe ingredients with food composition data," in *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, IEEE, vol. 1, 2015, pp. 330–336.

T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Knowledge extraction of evidence - based dietary recommendations," in *Proceedings of the Jožef Stefan International Postgraduate School Students Conference 2016*, Jožef Stefan International Postgraduate School, vol. 1, 2016, pp. 149–153.

T. Eftimov, B. Koroušić Seljak, and P. Korošec, "Grammar and dictionary based named-entity linking for knowledge extraction of evidence-based dietary recommendations," in *Proceedings of the 8th international Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, (IC3K 2016)*, vol. 1:KDIR, 2016, pp. 150–157.

T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Disadvantages of statistical comparison of stochastic optimization algorithms," *Proceedings of the Bioinspired Optimizaiton Methods and their Applications, BIOMA*, pp. 105–118, 2016.

T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Deep statistical comparison applied on quality indicators to compare multi-objective stochastic optimization algorithms," *Proceedings of the Third International Conference on Machine Learning, Optimization and Big Data, MOD 2017*, 2017.

T. Eftimov, G. Ispirova, P. Korošec, and B. Koroušić Seljak, "A semi-automatic system for classifying and describing foods according to FoodEx2," in *3rd IMEKO FOODS, Metrology promoting Standardization and Harmonization in Food and Nutrition*, 2017, pp. 56–59.

T. Eftimov, P. Korošec, and B. Koroušić Seljak, "The behavior of deep statistical comparison approach for different criteria of comparing distirbutions," in *Proceedings of the 9th International Conference on Computational Intellifence, (IJCCI 2017)*, vol. 1, 2017, pp. 73–82.

G. Ispirova, T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Mapping food composition data from various data sources to a domain-specific ontology," in *Proceedings of the 9th International Conference on Knowledge Engineering and Ontology Development, (KEOD 2017)*, vol. 2, 2017, pp. 203–210.

T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Comparing multi-objective optimization algorithms using an ensemble of quality indicators with deep statistical comparison approach," *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI 2017)*, pp. 2801–2809, 2017.

Conference Abstracts

T. Eftimov, S. Mezgec, P. Korošec, and B. Koroušić Seljak, "Mixed deep learning and natural language processing approach for food image detection, recognition and analysis to estimate nutritional values," in *Book of abstracts - Advancing Behaviour Change Science*, ISBNPA, 2017.

D. Potočnik, T. Eftimov, and N. Ogrinc, "Differentiation of Slovenian milk based on the content and carbon isotope composition of fatty acids," in *Book of abstracts: 5^{th} MS Food Day*, 2017, p. 246.

T. Eftimov, P. Korošec, B. Koroušić Seljak, M. Roe, A. Turrini, and P. Finglas, "Matching foods from EuroFIR databases to FoodEx by using a semi-automatic system for classifying and describing foods," in *Book of abstracts: 12th International Food Data Conference*, 2017.

Biography

Tome Eftimov was born on February 17, 1989 in Strumica, Macedonia. In 2011, he received a Bachelor of Science Engineering in the area of Informatics and Computer Engineering at the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Macedonia. In 2013, he received a Master of Science Engineering in the area of Computer networks and e-technologies at the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia. During the undergraduate studies he held a state scholarship for talented students awarded by the Ministry of Education and Science of Macedonia.

In 2011, he was teaching assistant in Probability and Statistics, Calculus 1, Calculus 2, Discrete Mathematics, Linear Algebra, and Numerical Methods at the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Macedonia. In 2012, he joined the Laboratory for Complex Systems and Networks at the Macedonian Academy of Sciences and Arts, where his research interests were in noncoherent communication, especially in applications of random matrix theory to the derivations of the performance limits in wireless communication. He was part of a DFG project, Noncoherent Communication.

In the Fall 2014, he started his PhD studies in Information and Communication Technologies at the Jožef Stefan International Postgraduate School in Ljubljana, Slovenia, under the supervision of Asst. Prof. Dr. Barbara Koroušić Seljak. He held a doctoral studies scholarship awarded by the project ISO-FOOD, ERA Chair for Isotope Techniques in Food Quality, Safety and Traceability, grant no. 621329 (2014-2019) and the Computer Systems Department at the Jožef Stefan Institute in Ljubljana, Slovenia. During his studies he took part in several EU funded projects ISO-FOOD (ERA Chair for Isotope Techniques in Food Quality, Safety and Traceability), RICHFIELDS (Research Infrastructure on Consumer Health and Food Intake for E-science with Linked Data Sharing), REFRESH (Resource Efficient Food and dRink for the Entire Supply cHain), and QuaLiFy (Quantify Life - Feed Yourself).

His research work is focused on developing new methodologies and tools for data management, exploration and exploitation. The main research areas are the Statistics, Natural Language Processing, Information Theory, Data Mining, and Evolutionary Algorithms. His current research is concerned with methods in the field of food- and nutrition-related data and knowledge management, especially focused on food matching, knowledge extraction of evidence-based dietary recommendations, harmonization of food- and nutrition-related data, and statistical analyses. He has published his work in several international journals and has presented it at several national and international conferences and workshops.