UNIVERSITÄT OSNABRÜCK

Interpretable machine learning for real-world applications

Dissertation
zur Erlangung des Grades
eines Doktors der Naturwissenschaften
eingereicht am Fachbereich Humanwissenschaften
der Universität Osnabrück

vorgelegt von
Olivera Stojanović

Osnabrück, März 2023

Gutachter:

1. Prof. Dr. Gordon Pipa
2. Prof. Dr. Peter König
3. PD. Dr. Georges Hattab

Standort: Osnabrück
Disputation: 11.10.2023

Interpretable machine learning for real-world applications

Dissertation
for a doctoral degree in natural sciences
submitted to the School of Human Sciences
of Osnabrück University

by
Olivera Stojanović

Osnabrück, March 2023

Supervisors:

1. Prof. Dr. Gordon Pipa
2. Prof. Dr. Peter König
3. PD. Dr. Georges Hattab

Location: Osnabrück
Disputation: 11.10.2023

The Three Laws of Robotics:

1: A robot may not injure a human being or, through inaction, allow a human being to come to harm;

2: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law;

3: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law;

The Zeroth Law: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Isaac Asimov, I, Robot

Contents

Abstract **v**

My scientific contributions **vii**

Preface **xi**

Acknowledgments **xiii**

1 Why machine learning models (often) fail **1**
1.1 Data-related challenges in machine learning . . . 2
1.1.1 Critical evaluation of the context of a dataset . . . 2
1.1.2 Challenges with benchmark datasets . . . 3
1.2 Model-related challenges in machine learning . . . 3
1.3 How interpretable machine learning can help . . . 6

2 What is interpretable machine learning **7**
2.1 Difference between explainable and interpretable machine learning . 9
2.2 Domain-specific model interpretability . . . 10
2.3 Challenges for interpretable machine learning . . . 11

3 Example 1: predicting epileptic seizures **13**

4 Predicting epileptic seizures using nonnegative matrix factorization **17**
4.1 Abstract . . . 18
4.2 Introduction . . . 18
4.3 Materials and Methods . . . 20
4.3.1 Data preparation . . . 20
Freiburg EPILEPSIAE dataset . . . 20
Epilepsyecosystem dataset . . . 20
4.3.2 Deriving time and frequency components . . . 21
4.3.3 Time-frequency decomposition . . . 22
4.3.4 Prediction and performance measures . . . 23
4.4 Results and discussion . . . 25
4.4.1 Interpretability of the model . . . 25
4.4.2 Predictive performance . . . 26
4.5 Conclusion . . . 28

5 Example 2: predicting infectious diseases 31
5.1 Properties of spatio-temporal data 32
5.2 Properties of public data 33
5.3 Modeling the spread of infectious diseases 33

6 A Bayesian Monte Carlo approach for predicting the spread of infectious diseases 35
6.1 Abstract 36
6.2 Introduction 36
6.3 Materials and methods 38
6.3.1 The BSTI Model 38
6.3.2 Monte Carlo sampling procedure 41
6.3.3 Interaction effects 41
6.3.4 Additional features 44
6.3.5 Predictive performance evaluation and model selection 45
6.3.6 The *hhh4* model reference implementation 45
6.4 Results and discussion 46
6.4.1 The inferred model 46
6.4.2 Predictive performance 49
6.4.3 Benefits of probabilistic modeling for epidemiology 52
6.4.4 Possible extensions 52
6.5 Conclusion 53

7 Example 3: predicting environmental variables 55

8 Bayesian hierarchical models can infer interpretable predictions of leaf area index from heterogeneous datasets 57
8.1 Abstract 58
8.2 Introduction 59
8.3 Methods 60
8.3.1 The dataset 60
8.3.2 Feature extraction from reflectance spectra with a spline basis 62
8.3.3 Bayesian Markov Chain Monte Carlo regression for predicting LAI 64
Model 1: A baseline model with pooled data 65
Model 2: A model with hierarchical bias 66
Model 3: Full hierarchical model 67
8.3.4 Model selection using Pareto-Smoothed Importance Sampling 68
8.3.5 Evaluation of feature importance 68
8.4 Results 69
8.4.1 Model predictions 69
8.4.2 Model comparison 69
8.4.3 An interpretable kernel function 71
8.4.4 Feature importance 73
8.5 Discussion 73

9 Conclusion **77**
9.1 Predicting epileptic seizures 78
9.2 Predicting infectious diseases 79
9.3 Predicting environmental variables 80
9.4 Broader conclusions from this thesis 80
9.5 Data science in the private sector 82
9.6 The future of interpretable machine learning 83
9.6.1 The importance of AI regulations 83
9.6.2 The importance of multidisciplinarity 84

A Supporting information: Predicting epileptic seizures using non-negative matrix factorization **87**
A.1 S1A Fig. 87
A.2 S2A Fig. 88
A.3 S3A Fig. 88
A.4 S4A Fig. 89
A.5 S5A Fig. 90

B Visualizing the spread of infectious diseases using public health data **91**

C Supporting information: A Bayesian Monte Carlo approach for predicting the spread of infectious diseases **93**
C.1 S1C Fig. 94
C.2 S2C Fig. 95
C.3 S3C Fig. 96
C.4 S4C Fig. 97
C.5 S5C Fig. 98
C.6 S6C Fig. 99
C.7 S7C Fig. 100
C.8 S8C Fig. 101
C.9 S9C Fig. 102
C.10 S10C Fig. 103

D Supplementary Material: Bayesian hierarchical models can infer interpretable predictions of leaf area index from heterogeneous datasets **105**
D.1 Figure S1D 106
D.2 Figure S2D 107
D.3 Figure S3D 108
D.4 Figure S4D 109
D.5 Figure S5D 110
D.6 Figure S6D 111
D.7 Figure S7D 112

Abstract

English

Recent severe failures of black box models in high stakes decisions have increased interest in interpretable machine learning. In this cumulative thesis, I discuss why black box machine learning models can fail and explain the potential of interpretable machine learning. After a general introduction into this topic, I present three examples of interpretable machine learning models that I developed for studies in different scientific fields: medicine, epidemiology, and remote sensing, which correspond to three publications that constitute the thesis. For each publication, I first explain the data context, the prediction task, why it is a challenging problem and how interpretable machine learning can help improve the outcome. Then, in each publication, I outline the methods, examine their performance, and discuss how interpretability adds to understanding the results and phenomena. The publications show that it is possible to design interpretable models that yield good predictions, but they also demonstrate that domain expertise and understanding of data context are crucial. The thesis concludes with an outlook on the future of interpretable machine learning. I argue that, especially when it comes to high stakes decisions, a better understanding of machine learning models will be crucial - also because new and future laws will increasingly regulate algorithmic decisions.

Deutsch

Ernsthafte Misserfolge von Black Box Modellen bei schwerwiegenden Entscheidungen haben in jüngster Zeit zu wachsendem Interesse an interpretierbarem maschinellem Lernen geführt. In dieser kumulativen Dissertation beginne ich mit einer Erläuterung, warum Black Box Modelle versagen können, und erkläre das Potenzial von interpretierbarem maschinellem Lernen. Anschließend stelle ich drei Beispiele für die Entwicklung interpretierbarer maschineller Lernmodelle in verschiedenen wissenschaftlichen Bereichen vor: Medizin, Epidemiologie und Fernerkundung, die drei Publikationen entsprechen die diese Arbeit ausmachen. Für jede Publikation erläutere ich zunächst den Datenkontext, die Problemstellung, warum es sich dabei um ein schwieriges Problem handelt und wie interpretierbares maschinelles Lernen verwendet werden kann, um das Ergebnis zu verbessern. Dann stelle ich in jeder Publikation die verwendeten Methoden vor, untersuche ihre Leistungsfähigkeit und erörtere, wie die Interpretierbarkeit zum Verständnis der Ergebnisse und Phänomene beiträgt. Die Publikationen zeigen zum einen, dass es möglich ist, interpretierbare Modelle zu entwerfen, die gute Vorhersagen liefern. Zum anderem, sie zeigen aber auch, dass Fachwissen und Verständnis des Datenkontextes entscheidend sind. Die Arbeit schließt mit einem Ausblick auf die Zukunft des interpretierbaren maschinellen Lernens ab. Ich argumentiere, dass insbesondere bei schwerwiegenden Entscheidungen ein besseres Verständnis von Modellen des maschinellen Lernens von entscheidender Bedeutung sein wird - auch weil neue und künftige Gesetze algorithmische Entscheidungen zunehmend stärker regulieren werden.

My scientific contributions

As part of this thesis, I contributed three peer-reviewed journal papers as the first author. Author contributions for each paper and of each author are listed here in order of authorship.

1. ***Stojanović O***, *Kuhlmann L, Pipa G (2020) Predicting epileptic seizures using nonnegative matrix factorization. PLOS ONE 15(2): e0228025. https://doi.org/10.1371/journal.pone.0228025*

   Olivera Stojanović
   Roles: Formal analysis, Project administration, Software, Validation, Visualization, Writing - original draft

   Levin Kuhlmann
   Roles: Data curation, Methodology, Supervision, Writing - review & editing

   Gordon Pipa
   Roles: Conceptualization, Data curation, Methodology, Supervision

2. ***Stojanović O***, *Leugering J, Pipa G, Ghozzi S, Ullrich A (2019) A Bayesian Monte Carlo approach for predicting the spread of infectious diseases. PLOS ONE 14(12): e0225838. https://doi.org/10.1371/journal.pone.0225838*

   Olivera Stojanović
   Contributed equally to this work with: Olivera Stojanovic, Johannes Leugering
   Roles: Conceptualization, Methodology, Project administration, Software, Visualization, Writing - original draft

   Johannes Leugering
   Contributed equally to this work with: Olivera Stojanović, Johannes Leugering
   Roles: Conceptualization, Methodology, Software, Visualization, Writing - original draft

Gordon Pipa
Roles: Conceptualization, Supervision

Stéphane Ghozzi
Roles: Data curation, Resources, Validation, Writing - review & editing

Alexander Ullrich
Roles: Data curation, Resources, Supervision, Validation, Writing - review & editing

3. ***Stojanović O**, Siegmann B, Jarmer T, Pipa G, Leugering J (2022) Bayesian Hierarchical Models can Infer Interpretable Predictions of Leaf Area Index From Heterogeneous Datasets. Frontiers in Environmental Science, 717. https://doi.org/10.3389/fenvs.2021.780814*

   Olivera Stojanović
   Roles: Conceptualization, Methodology, Project administration, Software, Visualization, Writing — original draft

   Bastian Siegmann
   Roles: Data curation, Conceptualization, Validation, Writing — review and editing

   Thomas Jarmer
   Roles: Data curation, Conceptualization, Validation, Writing — review and editing

   Gordon Pipa
   Roles: Supervision, Resources, Conceptualization, Validation, Writing — review and editing

   Johannes Leugering
   Roles: Conceptualization, Methodology, Project administration, Software, Visualization, Writing — original draft

I also presented these results in various talks and in a scientific poster:

1. *Visualizing the spread of infectious diseases using public health data*, doi: `10.13140/RG.2.2.29991.50083`

All used code is available in the following code repositories:

1. `https://github.com/ostojanovic/seizure_prediction`

2. `https://github.com/ostojanovic/BSTIM`
3. `https://github.com/ostojanovic/visualizing_the_spread_of_infectious_diseases_using_public_health_data`
4. `https://github.com/ostojanovic/bayesian_lai`

Preface

What is this thesis about?

In the first two chapters, I write about the common challenges in machine learning that stem from the complexity of the data context and the use of interpretable machine learning to mitigate them. Then I show examples of such models in three scientific fields:

- medicine - for which I developed a method for prediction of epileptic seizures,
- epidemiology - for which I developed a method for prediction of the spread of infectious diseases and
- remote sensing - for which I developed a method for predicting the leaf area index.

The thesis is organized as follows: for each of the models, an overview of the prediction task and challenges in the scientific field are given in one chapter, and the developed model is shown and analyzed in the next chapter. In chapters 3 and 4, I describe our model for the prediction of epileptic seizures. In chapters 5 and 6, I talk about the model for the prediction of infectious diseases, and in chapters 7 and 8, I talk about the model for prediction of the leaf area index. I conclude the thesis with an overview of the developed models and an outlook on the future of interpretable machine learning.

How did this thesis come to be?

I started my Ph.D. studies at the Institute of Cognitive Science after finishing my bachelor's studies in Physics (focusing on Medical Physics) and master's studies in Electrical and Computer Engineering at the University of Novi Sad in Serbia. Since I had previously worked with data analysis of electroencephalographic (EEG) data, my initial interest was medical data in general and EEG data in particular. Following this direction, I wrote the first paper, *"Predicting epileptic seizures using nonnegative matrix factorization"*. During this project, we collaborated with Dr. Levin Kuhlmann from Monash University in Melbourne, Australia. Working on this project showed me the complexities of real-world data and the importance of data context. I got interested in interpretability of machine learning models, and I wanted to expand my research into time series prediction.

Following these interests, I wrote my second paper, *“A Bayesian Monte Carlo approach for predicting the spread of infectious diseases”*. The paper is a result of the collaboration with Dr. Alexander Ullrich and Dr. Stéphane Ghozzi from the Robert Koch Institute in Berlin. As a part of the project, I spent two months at the Signale Group of the Robert Koch Institute, working on implementing the proposed model. During this time, I learned a lot about Bayesian methods and their potential for developing interpretable machine learning models. Further, I studied how interpretable models help communicate scientific results to a broader audience. As a part of the project, I presented a poster at the EU Data Viz conference in 2019 entitled *“Visualizing the spread of infectious diseases using public health data”*. The poster presented our developed methods, how they could be used to serve citizens through data visualization and to communicate epidemiological data to the public.

Continuing my interest in the Bayesian approach, I wrote the third paper *“Bayesian hierarchical models can infer interpretable predictions of leaf area index from heterogeneous datasets”*. It explores the possibilities of Bayesian hierarchical models for developing interpretable machine learning models for application in remote sensing and, more broadly, environmental sciences. In this project, we collaborated with Dr. Bastian Siegmann from Jülich Research Centre and Dr. Thomas Jarmer from the Institute of Computer Science in Osnabrück. Here, I used Bayesian hierarchical models to deal with challenges similar to those I encountered in previous projects: making predictions from limited and heterogenous datasets.

Currently, I work as a data scientist in the private sector. My daily work includes analyzing various datasets on different spatial levels (e.g., estimating purchasing power for different administrative levels in Germany or other countries). In my work, analyzing heterogeneous datasets and understanding the data context is crucial for interpreting models based on such data and for further communication with clients.

Olivera Stojanović
Nürnberg, March 2023.

Acknowledgments

This thesis is a product of three projects from very different domains, but all of them show the potential of interpretable machine learning. During my Ph.D., I was given a lot of freedom to explore my scientific interests and focus on interpretable machine learning models before they became part of a larger scientific discourse.

I would like to thank my supervisor, Prof. Gordon Pipa, for allowing me to work on such diverse projects and for trusting me with the topic of my thesis.

I would further like to thank the Signale Group of the Robert Koch Institute, especially Dr. Alexander Ullrich and Dr. Stéphane Ghozzi, for their collaboration on the project of predicting the infectious diseases, their invaluable domain expertise and for giving me a chance to spend the time at the Robert Koch Institute.

I would also like to thank our other collaborators, Dr. Levin Kuhlmann from Monash University in Melbourne, Australia, for his help and expertise in predicting epileptic seizures, and Dr. Bastian Siegmann from Jülich Research Centre and Dr. Thomas Jarmer from the Institute of Computer Science in Osnabrück, for providing data and their helpful feedback on the project of predicting the leaf area index.

Last but not least, I would like to thank my family and my fiancée Johannes for their continuous support throughout these years.

Chapter 1

Why machine learning models (often) fail

At the beginning of the coronavirus pandemic, there were high expectations for artificial intelligence (AI) and machine learning to help curb the pandemic. We were hoping for AI to help us stop the spread of coronavirus, trace contacts, diagnose COVID-19[1], and provide up-to-date information. There were reasons for optimism, since this is the first pandemic in which we have supercomputers and powerful machine learning models, we know how to work with big datasets, and we are more connected than ever.

Three years after the beginning of the pandemic, these expectations are still not met. Although a lot of prediction models were developed, most of them underperformed, or their predictive performance is still not suitable for clinical use. [2, 3, 4, 5] We got contact tracing apps, but some of them infringe on data privacy for purposes that users did not initially agree to (e.g., police in Germany requesting data from the "Luca" contact tracing app [6]). This erodes trust in the apps, and makes contact tracing more difficult.

How did this happen? Why did AI fail to live up to its potential in this instance? The Alan Turing Institute lists four main points as sources of problems: incomplete or poor quality data, automated discrimination, human errors, and challenges in communication between researchers, policy makers, and the public. [7] Data was collected in real-time across many different institutions and countries, all of which have different standards, and the datasets, in the beginning, were not large. It is generally hard to predict the spread of the virus from small, not standardized, heterogeneous datasets, where the chance for human errors is high.

Further, available datasets in medicine usually reflect historical discrimination and biases, such as unequal and often unethical treatment of marginalized groups without their consent (e.g., Tuskegee Syphilis Study [8], The Puerto Rico Pill Trials [9]). This leads to incomplete medical data and distrust of the affected groups toward the whole healthcare system. [10, 11] When we train machine learning models on such data, algorithms will incorporate these inequalities, which

[1] According to the World Health Organization, the official name of the new virus is *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2), and the disease it causes is COVID-19. [1] I will in this thesis refer to the virus shortly as *coronavirus*.

also happened during the coronavirus pandemic. Many predictions and recommended decisions of hospital management algorithms led to discriminatory treatment of patients. [3] To address this issue, we need to inspect the background and context of data well before developing a model, and we need to make sure that we understand what the model learns.

The problems pointed out by The Alan Turing Institute are the same ones that often happen in machine learning projects, and this is a good chance to learn how to avoid them in the future. They can be loosely separated into data- and model-related challenges.

1.1 Data-related challenges in machine learning

1.1.1 Critical evaluation of the context of a dataset

The biggest challenge in machine learning is often thought to be developing algorithms, but creating, curating, and standardizing large datasets can take more time and resources. We often need to collect data from different sources, use different equipment, etc., and the chance of human error is high. Except for benchmark datasets like ImageNet [12], MNIST [13] or CIFAR-10 [14], many available datasets were not collected with a specific machine learning purpose in mind. [15, 11] Such datasets sometimes don't have crucial information about data collection processes, which makes further analysis difficult.

Further, data depends on the environment in which it was created. Donna Haraway introduced the concept of *"Situated knowledges"*, a term that describes embodied objectivity of scientific knowledge. [16] She argues that there is no "view from above" where it is possible to see everything, but rather that knowledge always comes with a certain perspective. [16] Putting this into a data science perspective, we can talk about situated data systems consisting of data and their context. [11] To fully understand data, one has to understand the circumstances in which they were created. This means the data context is also a part of the model. If the background of a dataset is not carefully examined before training a model, there is a risk of creating a model that will repeat the same errors or biases encoded in the dataset. [17, 18, 15, 19]

For example, in the context of the coronavirus pandemic, various countries released data on the number of cases, confirmed deaths, conducted tests, etc. When working with such data, we have to check testing conditions because if fewer tests are carried out, there may be many more cases than officially reported. In such a case, looking at the excess mortality rate might be more interesting because it shows a better picture of the situation. However, there might be different definitions for what counts as death from COVID-19. [20] Finally, it is crucial to consider whether some data might be missing or or not fully collected, e.g., the number of confirmed cases might be misrepresented or underreported. [21] This expert knowledge is important for model development, comparison among countries, or communication with the public. [22, 23, 11, 24, 25]

1.1.2 Challenges with benchmark datasets

In machine learning, the quality of a method is typically evaluated by its performance on a given benchmark set. However, benchmark datasets are not equally standardized across research fields. [26] Traditionally, the first big curated datasets have been created in computer vision, where they are still most prevalent [27, 12, 13]. Collecting data in fields such as medicine, epidemiology, remote sensing, psychology, etc., is more challenging. [28] For example, it takes a long time and many resources in a clinical setting to collect, standardize, and publish data. [29] This process requires medical and data experts as well as legal and ethical expertise. As a result, datasets in medicine can be heterogeneous and imbalanced, and models trained on such data can achieve misleading high accuracy by always predicting the majority class. [29]

However, even if an algorithm has a high prediction accuracy on a benchmark dataset, there is still a possibility that the method is learning how to “win” the specific benchmark rather than learning general patterns in data. [30] This happens in fields with very few benchmark (or regular) datasets. For example, in epileptic seizure prediction, the EPILEPSIAE dataset [31] has long been used as one of the benchmarks for new seizure prediction algorithms. Since data in the field is limited, the performance of many proposed algorithms is only calculated on the EPILEPSIAE dataset. [32] This is a problem for application in the real world, e.g., for predicting epileptic seizures in real-time, because such algorithms will not be able to generalize well.

1.2 Model-related challenges in machine learning

Many successful machine learning methods are black box algorithms, i.e., models in which humans don’t understand how or what they learn. There are two types of black box machine learning models: *complicated*, i.e., models too complex for humans to comprehend, and *proprietary*, i.e., models to which the public doesn’t have access, often due to trade secrets. [33] Interpretability is not needed if the consequences of imprecise predictions are not critical, like in ad servers. Also, tasks that are well studied and validated in real applications, with enough practical experience, such as models for optical character recognition to extract addresses from envelopes, do not require an interpretable model. [34, 35]

However, all machine learning models perform better during development and on the same benchmark datasets on which they were trained than in practical situations on new datasets. This difference is usually acceptable, but problems arise when predictions are used to justify high stakes decisions. [36, 37, 38, 39] In such situations, a better understanding of what the model in question learns and how it combines features for predictions is needed, but black box models are not designed for this. [33]

For example, during wildfires in California in 2018, Google was using a black box machine learning model made by BreezoMeter to predict air quality. Usually, the algorithm has a prediction accuracy of 98.4%. [40] However, during the wildfires, the model predicted the air quality was “good - ideal air quality for outdoor activities" when in reality, ash was in the air and was dangerous to breathe. [41]

In another example, the government of the United Kingdom (UK) decided to use an algorithm to assign final grades for A-levels in the summer of 2020 since schools were closed in the previous term because of the coronavirus pandemic. The algorithm reduced the grades of students from more impoverished schools and upgraded the grades of students from private schools, going against already given school grades. [42, 43] After the backlash from students and the public, the algorithm had to be abandoned, and teachers instead gave the grades manually. [42]

These models are proprietary, and because of this, it is impossible to know whether they are complicated or not. We can only speculate about the potential causes of these predictions, such as an overfitted model for risk assessment of air quality or the impossibility of encoding complex variables (e.g., capability and work of students, teachers' capacity and assessments, quality of schools, etc.) into a model without biases.[2]

Similar examples where black box models for high stakes decisions led to unintended consequences are:

1. Schufa Holding AG (German for *Schutzgemeinschaft für allgemeine Kreditsicherung*, in English *General Credit Protection Agency*) is a private company which provides credit ratings in Germany, commonly known as SCHUFA scores. [44] The scores are used for various situations, like renting a place to live, credit card applications, making a new Internet contract, etc. [45, 46] The algorithm for calculating SCHUFA scores is a proprietary black box. In 2018, non-profit organizations AlgorithmWatch [45] and Open Knowledge Foundation Germany [47] started the OpenSchufa project. [48] They asked consumers to upload SCHUFA anonymously to a common database for research into the algorithm. Investigative journalists from *Der Spiegel* and *Bayerischer Rundfunk* evaluated this data. An investigation concluded that most of the data used to calculate scores are based on address, age, and gender. [45, 48] They showed that some people got negative scores even though they didn't have negative indicators, such as debts. [46] Further, they showed that younger people tend to get a lower score than older people with similar characteristics, which might be a reflection of the demographic structure of Germany. [49] If this is the case, such scoring system would not add new information. It is important to note that the OpenSCHUFA database contains just a small part of a large database, and it might be distorted (i.e., some groups might be overrepresented), which would influence the results of the investigation. As long as SCHUFA is a proprietary black box algorithm, it will be hard to learn how SCHUFA scores are calculated and whether these assessments are unbiased. So far, the company has published SCHUFA Simulator, a set of questions one can answer to see what influences the SCHUFA score. [50] But

[2] There is also the question of whether or not it is ethical to develop such an algorithm in the first place because it takes away the agency from students to show their capabilities and assigns them grades for which they can not trace how they were given.

the score it provides is only close to, but not exactly equal to the SCHUFA score, and the simulator does not show how these factors are combined.

2. In recent years, an increasing number of emotional recognition systems, an extension of facial recognition systems, have been released. [51] These systems usually use micro-movements of muscles, eye movements, facial expressions, and changes in the voice to estimate the emotional state of the subject. Creators of this technology claim that their algorithms can understand human feelings and intentions, from which they can predict future behavior.[51] For example, software for the detection of emotions is being used in high schools in Hong Kong to detect lapses in attention of students during distance learning. [52] Some companies already use such systems for assessments of candidates during job interviews or for customers' reactions on advertising. [53, 51] Frontex, the European Border and Coast Guard Agency, is currently investigating systems for automated emotional recognition, which are intended to be employed at the Schengen borders. [54] Such algorithms are proprietary, and possibly complicated black box models. Research shows that while such algorithms can decode facial expressions, they fail to decode and predict feelings or intentions. [55] How people express themselves emotionally varies a lot and is highly contextual. Analyzing facial expressions using only black box models is not a reliable predictor for emotions or intentions. [56, 57] Further, emotional expressions vary among cultures, and it is not clear that an AI system would be able to pick up on these differences, if not explicitly designed to do so. [58, 59]

3. Lately, there has been an increasing interest in the application of AI for the selection of embryos in *in vitro fertilization* (IVF). The AI has the potential to evaluate steps in the IVF process better and to make them more reproducible and faster. [60] Many of the developed algorithms are proprietary or complicated black box algorithms, or in some cases, algorithms that combine interpretable features in an uninterpretable way. [61] This leads to many ethical and technical concerns. The main issue is that no black box model has been evaluated using randomized controlled trials, which are necessary for evaluating clinical trials. [61, 62] Further, most of these black box models were designed to select embryos with a higher chance of implantation. It might happen that some disadvantageous traits, like an increased risk of cancer, correlate with higher chances of implantation. Whether we should choose an embryo with the highest chance of implantation or an embryo with the best chance of the best life is an ongoing philosophical debate [63, 64], for which parents and clinicians have to be involved in the decision. However, black box models do not allow for shared decision-making with patients. Shared decision-making in medical systems helps domain experts correct mistakes or unwanted decisions made by the model and increases patients' trust. [65, 66, 67] A more in-depth discussion on this issue and potential solutions involving interpretable machine learning are given in [61].

In all of the cases, if more interpretable methods had been used instead of black box machine learning models, it would have been easier to notice what they learned and correct mistakes in the pipeline before their widespread use (or abandonment in the case of the grading algorithm).

1.3 How interpretable machine learning can help

Because of these challenges, there has been an increased interest in interpretable machine learning. Learning interpretable features and creating interpretable models has a long tradition in statistics (e.g., linear regression, generalized additive models, elastic net, etc.). The idea behind these methods is to make certain assumptions about a probability distribution explicit or to restrict model complexity, which leads to more inherently interpretable models. The success of black box machine learning models has led many to believe that interpretable machine learning methods are “old-fashioned” and incompatible with the demands of big data. [68] However, cases such BreezoMeter air quality predictions or the UK grading algorithm show that we also need interpretable models. Interpretability is often posed opposite to accuracy, but this is not necessarily the case. Interpretability might help achieve higher accuracy because it makes it possible to understand the model better and troubleshoot it. [69]

In the last couple of years, more work has been done to define interpretability [70, 71], to understand the strengths and weaknesses of interpretable methods [34, 72, 35, 73, 71], and to evaluate them. Interpretable machine learning has reached “the first state of readiness”, i.e., with the development of interpretable methods, more software for implementing them has been written in the public and private sectors. [74, 75, 76, 77, 78, 79, 80]

Interpretable machine learning benefits domain experts, decision-makers, and the general public. Domain experts are often interested in understanding the dynamics of the process, and it has been shown that people who work with interpretable algorithms have more trust in the method and their team members. [81] Providing interpretable machine learning predictions also makes decision-makers less skeptical of machine learning methods and more willing to accept them. [82] Finally, humans trust decisions and systems more, if they think they understand how they work. [33, 83]

Chapter 2

What is interpretable machine learning

There are various definitions of interpretability in machine learning, depending on whether the focus is on understanding the given model or understanding the predictions it makes. Some of them are:

> *Interpretability is the degree to which a human can understand the cause of a decision.* [84]

> *Interpretability is the degree to which a human can consistently predict the model's result.* [85]

> *We define interpretable machine learning as the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data. Here, we view knowledge as relevant if it provides insight for a particular audience into a chosen domain problem.* [71]

> *An interpretable machine learning model obeys a domain-specific set of constraints to allow it (or its predictions, or the data) to be more easily understood by humans. These constraints can differ dramatically depending on the domain.* [69]

Although defitions are diverse, we tend to consider models interpretable if they share some common properties [86]:

- **trust** - models that humans feel comfortable using and giving control to them,
- **causality** - models which learn causal relationships between variables,
- **transferability** - models that allow us to transfer learned skills to a new environment,
- **informativeness** - models that provide useful information about the real world and

- **fair and ethical decision-making** - models that make predictions that align with human values.

Because of the diverse definitions and properties of interpretable models, there are different ways to design them. One way is to focus on algorithm transparency, a property that describes how an algorithm learns relationships between variables in general. [35] Here the focus is not on interpreting the model for given data or predictions but on the theoretical understanding of algorithms. Such algorithms, because of their transparency, are used in decision-making systems. Because the task in decision-making systems is to choose the best option among several courses of action, it is helpful to understand how an algorithm works internally. Examples of highly transparent algorithms are the least squares method, rule lists, decision trees, etc. [35]

Another way to design interpretable machine learning algorithms is to focus on local and global model interpretability, which are not properties of algorithms themselves. [35] Rather, they refer to a set of methods and ways of combining different approaches to interpret why models make specific predictions for sample datapoints (local model interpretability) or how the trained model makes predictions globally by learning relationships between features, learned components, and their contribution to predictions. [71, 35] However, this is challenging in models with lots of parameters and high-dimensional data. In such cases it is often only possible to interpret how parts of the model affect predictions (global interpretability on a modular level). [35]

Techniques for achieving local model interpretability are usually model-agnostic methods we use on a subset of data. For example, local interpretable model-agnostic explanations (LIME) are often used with support vector machines (SVM), neural networks, random forests, etc., and work well for tabular data, text, and images. [70, 87, 88] The procedure works as follows: we first select a datapoint of interest (a point for which we want to have an explanation of the prediction). We then perturb the input data and get predictions with the original (not interpretable) model. LIME then trains an interpretable model on the perturbed input, weighted by how close the new samples are to the datapoint of interest. In the last step, we explain the prediction by interpreting the local model.[1]

Similarly, model-agnostic methods are also used to achieve global model interpretability. The difference here is that we are not interested in why a model makes certain predictions for specific datapoints, but we want to know how parts of the model influence predictions. For example, we often want to estimate how certain features influence predictions. Permutation feature importance is a model-agnostic method of measuring the decrease or increase in loss of the whole model when we permute the observations of a feature of interest in the test (train) data. [89] The method might seem similar to LIME, but the difference is that we are not interested in replacing a part of a model with an interpretable surrogate model. Instead, permutation feature importance uses the same model but is tested (trained) on permuted data for each feature. In this way, the method breaks

[1] It should be clarified that this is different from training linear models to explain black box models since we here tweak specific features locally to understand better how they impact predictions.

dependence between features and we can observe how the loss of the model changes. If loss increases, it indicates that the permuted the feature is important for predictions.

2.1 Difference between explainable and interpretable machine learning

Interpretable machine learning is not the same as explainable artificial intelligence (XAI). Explainable AI focuses on providing post-hoc explanations of predictions, independent of whether the employed model is interpretable or not. [71] To develop an explainable model, one usually trains a new model or parts of the model on the same data on which the original black box model was trained. The explainable model should have similar predictions as the black box model, but the features that these two models learn and how they are combined can be different. [33] Essentially, the goal of explainable models is not to understand what the original model *does*, but to *resemble* predictions of the original model with a surrogate model that we understand. [33]

To draw this distinction more clearly, let's look at a black box model for the prediction of recidivism, proposed explanatory models, and why the explanatory models still can't fully interpret the original model. In 2000, software for predicting recidivism, known as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), was introduced in several American states. The purpose of the algorithm is to provide a fast, objective, and data-based assessment of the potential future behavior of defendants. [90] COMPAS uses 137 questions to assess whether an arrested person is at risk of committing a crime again in the near future. Questions range from a person's residence, workplace, and education level to whether anyone in their family committed a crime. Judges then get the score and decide whether and how to to use it in sentencing. [90]

After some time, it was noticed that the algorithm assigns higher risk scores to black Americans who committed minor crimes and lower risk scores to white Americans who committed major crimes. [90] However, the severity of the committed crime is not a feature in calculating the COMPAS score, and there are no explicit questions about the race of defendants. [90, 91, 92, 93] Since the algorithm is proprietary, its implementation details and complexity are not provided to the public, and it is unclear whether the algorithm predicts recidivism in the manner the creators or the public wanted to. [2] [90]

ProPublica, a non-profit organization for investigative journalism, first reported on COMPAS and made an explanatory model. The investigation concluded that the algorithm uses race as a defining feature. [91] This assessment is based on a linear model they developed to explain COMPAS' predictions. [91] The problem is that the original algorithm seems to be nonlinear, whereas ProPublica's model is linear. [95] Therefore, even though ProPublica's model can match COMPAS' predictions well, the linear explanatory model cannot show conclusively what the

[2] Similar to the UK grading algorithm, there is also a question of whether or not it is ethical to develop such models in the first place because defendants have not committed new crimes *yet*. See [94] for more discussion on this topic.

black box model intrinsically learns. There are no explicit questions about race, but other questions ask for information correlated with race (i.e., age, residence, income, etc.). Relations between these variables and crime are complex and highly contextual, which the linear model cannot capture. [33, 95] Essentially, we can learn from the explanatory model that race matters, but we can't learn *how* it matters, i.e., we don't know how COMPAS combines features for the final score. Instead of an explanatory model, an interpretable model might be more useful in this case.

One suggestion that could solve an issue with COMPAS are computer-aided exploration techniques. They consist of a model and a design interface where domain experts can change the model and see how their changes influence predictions. [96, 97, 69, 98] This would include humans in the decision-making, and they would see better how the model combines the variables instead of purely relying on COMPAS scores. The inability to assess how predictions are calculated makes it hard to combine predictions with knowledge outside the database, i.e., evidence for or against defendants. Judges are supposed to use the score for their assessments merely as an additional piece of information. Still, because they do not know how the model combines different factors, it is impossible to estimate whether their final assessment would, in reality, decrease the risk of recidivism or not. If there were an option to see how the model works intrinsically, lawyers and defendants would have the option to prepare for the trial, and there might be less bias in the final sentencing.

A theoretical case can be made that it should often be possible to construct inherently interpretable models. If many machine learning models can perform similarly on the same dataset, they should include some inherently interpretable models. Rudin et al. formalize this by the *Rashomon set of good models* [99, 89], which is defined as a set of high-performing models on the same dataset. They argue that if the Rashomon set for a given dataset is large, there must be one model which is simpler and interpretable, which should be taken in high-risk decisions instead of the black box model. [69, 100]

2.2 Domain-specific model interpretability

To develop an interpretable model, it is necessary to consider what interpretability in a given scientific field means. Domain experts often develop interpretable models by incorporating well-studied features or adding domain-related constraints. Because data are often high-dimensional, interpretable features are often constructed using unsupervised learning or dimensionality reduction. For example, nonnegative matrix factorization (NMF) is a decomposition method that decomposes a dataset into a (nonnegative) combination of (nonnegative) components or "patterns". [101] From the resulting components, domain experts can choose well-understood features while excluding others, such as noise. The data can then be represented by a specific combination of these features, which reduces the dimensionality considerably. [102] This is useful in fields like physics, bioinformatics, astronomy, etc. Other constraints like sparsity or monotonicity of variables can be enforced on models with many parameters. For example, if a model has many parameters, not all will contribute equally to predictions. And if we know that a model should learn only a few features but don't know which ones, we can use lasso regularization [103] to set less important parameters to zero, which helps identify relevant features for prediction.

2.3 Challenges for interpretable machine learning

Rudin et al. list ten grand challenges for the field of interpretable machine learning [69]:

- optimization of sparse models (for tabular data) and scoring systems,
- challenges involving generalized additive models (e.g., how to use them to troubleshoot complex datasets),
- challenges involving extending case-based reasoning to more complex datasets,
- supervised and unsupervised disentanglement of concepts in neural networks,
- dimension reduction for data visualization,
- incorporating physics or causal constraints in a model,
- understanding, exploring, and measuring the Rashomon set of accurate predictive models and
- interpretable reinforcement learning.

I encountered some of these challenges in my projects, like incorporating physics or causal constraints in a model and dimensionality reduction, as well as working with limited, heterogenous datasets with specific data context and field-related constraints. In the following chapters, I will show how we designed models in three scientific fields: medicine (prediction of epileptic seizures), epidemiology (prediction of infectious diseases), and environmental sciences (prediction of leaf area index). The models we created achieve local model interpretability (prediction of epileptic seizures) and global interpretability on a modular level (prediction of infectious diseases and prediction of leaf area index).

Chapter 3

Example 1: predicting epileptic seizures

Epilepsy is a neurological disorder that by current estimates affects 1% of the global population. [104] It is characterized by sudden seizures varying in length and intensity, which can lead to physical injuries, stress, and even death. [105] EEG is a powerful tool for diagnostics and monitoring epileptic seizures. A better understanding of epileptic seizures would help to anticipate, intervene, and even prevent them. In particular, AI-based seizure prediction has a lot of potential, and may help to develop closed-loop EEG devices for seizure prediction and intervention (e.g., neurostimulation, contacting caretakers). [106] These systems provide real-time monitoring of patients, giving them time to prepare, ease the seizure, and help understand epilepsy by collecting more data. [107, 32, 108, 109] Algorithms for closed-loop settings must be fast, easy to use, and not require a lot of memory storage. Despite many proposed seizure prediction models, due to technical difficulties, there have been few studies on implementing a closed-loop device in the human brain. [110, 111]

Seizure prediction algorithms focus on classifying between "normal" EEG and the "abnormal" EEG signal minutes before a seizure. Two goals of seizure prediction models are: to warn patients about upcoming seizures and to help medical professionals and domain experts understand epilepsy. Patients need an accurate algorithm that doesn't miss seizures (low rate of false negatives). Doctors and researchers are interested in learning about disease dynamics, understanding the model's behavior, and using the prediction model to help patients during or after medical procedures. Therefore, predicting epileptic seizures can inform high stakes decisions. In this setting, interpretability is important since the consequences of bad prediction can lead to wrong treatments and distrust of patients. Also, when using seizure prediction methods in a clinical setting, there is a question of liability for the decisions made based on their predictions.

In an EEG signal, seizures are distinguished by sudden hyper synchronization. [105] There are four main segments of EEG signals in epilepsy which correspond to four states patients can be in:

- **interictal** - a state between seizures,

- **preictal** - a state before a seizure,
- **ictal** - a state during a seizure,
- **postictal** - a state after a seizure.

There are quite a few data- and model-related challenges when designing seizure prediction models. First, collecting quality data on epileptic seizures in a clinical setting is challenging, expensive, and can take a long time. Not every hospital has the equipment or resources to do it, and there are ethical considerations. Seizures are rare events, and patients often have to undergo surgery, but we need to collect a large amount of data containing both interictal and preictal states. The EEG measurements have to be long enough to get sufficient data, but not so long that the conditions of patients worsen before the surgery. This leads to heterogeneous and imbalanced data, with more interictal recordings than preictal or ictal states collected per patient (which often have different diagnoses), using different EEG settings. [112, 32, 113] Developed models for seizure predictions must be patient-specific because of the limitations of the data collection process. This further limits the amount of available data and makes it hard to generalize the dynamics of epilepsy. Seizure prediction models also have to be able to deal with imbalanced datasets while preventing overfitting. Because of these challenges, only a few big datasets of EEG recordings of epileptic seizures are available for broader use. The two biggest and most often used ones are: the EPILEPSIAE dataset [31] and the Epilepsyecosystem dataset [114]. Since most methods are evaluated on one of the two available datasets, there is a danger of overgeneralizing the datasets and not learning useful features for potential use in a closed-loop EEG device. [32]

Seizure prediction methods usually combine features derived from EEG signals (e.g., spectral features, EEG patterns, etc.) and machine learning algorithms like SVM, decision trees, logistic regression, methods for time-series analysis, or neural networks. [107, 115, 32, 109, 116] Although there is a recent increase in interest in interpretable machine learning models for seizure prediction, most combine black box models and a post-hoc explanation method [117, 118, 119, 120, 121, 122] rather than designing inherently interpretable models. [123, 124] Here I will show how we developed a locally interpretable seizure prediction method and summarize the highlights of our paper.

We wanted to design an interpretable model and started by designing interpretable features of preictal and interictal states, which we used to classify between the states. To do so, we extracted the time and frequency components of intracranial EEG signals for each channel of each state for each patient using nonnegative matrix factorization, which, as mentioned earlier, is an interpretable and transparent decomposition method. [125] The components capture the dominant information from power spectra and detect structure in preictal states, which we use for classification. Learned time and frequency components are also informative for domain experts because they can be related to well-understood physical phenomena. We combined both major datasets, EPILEPSIAE and Epilepsyecosystem, to ensure that our model is robust.

We combined a linear support vector machine with L1 regularization for classification. Since SVM is not an algorithmically transparent classifier, we compensated for this by using L1 regularization to select informative EEG channels and weigh their contribution to the prediction. Combining these methods makes it possible to look at individual measurements, their NMF components, and the learned weights of L1 regularization. This way, we can see why the classifier assigns one of the two classes to a particular measurement and achieve local model interpretability. As the last step, we applied the synthetic minority over-sampling technique (SMOTE) to mitigate the class imbalance of interictal over preictal states. Our method produces good results and is computationally inexpensive, which could lend itself to an application in a closed-loop setting.

Chapter 4

Predicting epileptic seizures using nonnegative matrix factorization

Contributions

This chapter was first published as: *Stojanović O, Kuhlmann L, Pipa G (2020) Predicting epileptic seizures using nonnegative matrix factorization. PLOS ONE 15(2): e0228025. https://doi.org/10.1371/journal.pone.0228025*

Olivera Stojanović
Roles: Formal analysis, Project administration, Software, Validation, Visualization, Writing - original draft
Affiliation: Department of Neuroinformatics, Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany

Levin Kuhlmann
Roles: Data curation, Methodology, Supervision, Writing - review & editing
Affiliation: Data Science and AI Group, Faculty of Information Technology, Monash University, Clayton, Victoria, Australia

Gordon Pipa
Roles: Conceptualization, Data curation, Methodology, Supervision
Affiliation: Department of Neuroinformatics, Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany

Data Availability

Two datasets are used for evaluation of the method: EPILEPSIAE and Epilepsyecosystem. The EPILEPSIAE database is not publicly available. The European Epilepsy Database was developed in the EU-founded FP7 eHealth project EPILEPSIAE (Grant 211713). The database is owned by a third party and is commercially accessible for users who apply to the research groups in charge (http://epilepsy-database.eu). The Epilepsyecosystem database is publicly available. The Epilepsyecosystemdataset is free of charge and available upon registration at: https://www.epilepsyecosystem.org/register. The instructions for downloading the dataset can be found at: https://www.epilepsyecosystem.org/howitworks. For all inquires and questions, please refer to: levin.kuhlmann@monash.edu The code for this paper can be found at: https://github.com/ostojanovic/seizure_prediction.

Funding

This study was funded by the National Health and Medical Research Council (GNT1160815) to Dr Levin Kuhlmann. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors have declared that no competing interests exist.

4.1 Abstract

This paper presents a procedure for the patient-specific prediction of epileptic seizures. To this end, a combination of nonnegative matrix factorization (NMF) and smooth basis functions with robust regression is applied to power spectra of intracranial electroencephalographic (iEEG) signals. The resulting time and frequency components capture the dominant information from power spectra, while removing outliers and noise. This makes it possible to detect structure in preictal states, which is used for classification. Linear support vector machines (SVM) with L1 regularization are used to select and weigh the contributions from different number of not equally informative channels among patients. Due to class imbalance in data, synthetic minority over-sampling technique (SMOTE) is applied. The resulting method yields a computationally and conceptually simple, interpretable model of EEG signals of preictal and interictal states, which shows a good performance for the task of seizure prediction on two datasets (the EPILEPSIAE and on the public Epilepsyecosystem dataset).

4.2 Introduction

The ability to predict epileptic seizures provides an opportunity to intervene in order to attenuate their effects, or if possible prevent them. In this study we focus on EEG manifestations of seizures, which are characterized by sudden hypersynchronization of neurons and last from seconds to minutes. [105] Recently published studies on seizure prediction use a wide variety of approaches, from time series analysis (e.g. phase synchronization [126] or bivariate phase synchrony [127]) and spectral features of EEG signals [115, 128] to physiological models of neural activity (e.g. neural mass models[129]) or circadian models [130]. We focus on spectral measures of EEG signals since they have been successfully used as features for seizure prediction, and are easily interpretable. [107, 115, 32]

In the field of seizure prediction there are certain conceptional, computational and data-related challenges. First, using a large number of features for prediction makes it difficult to interpret their individual contribution. [32] Secondly, the algorithms for seizure prediction in a clinical setting need to be computationally efficient. Due to hardware constraints, this applies to closed-loop EEG devices for seizure prediction and intervention in particular, which have been a recent focus in the field. [107, 32, 108, 109] Finally, data encountered in the field of seizure prediction can be high dimensional and heterogeneous (e.g. recorded using many different channels and types of measurements in addition to EEG, like ECG, EOG etc), yet suffer from class imbalance (patients spend more time in interictal than in preictal states) and limited in the number of labeled samples. This is particularly challenging for the design of a patient-specific model.

In this study we address these issues by developing an easy-to-use, computationally efficient method for patient-specific seizure prediction. In order to achieve that, we extract a small set of interpretable features from power spectra that distinguish a baseline (interictal) EEG activity from a state leading up to a seizure (preictal state). Interictal states are regular brain activity between seizures, which can sometimes be interrupted with interictal spiking. [105, 131] Since seizures are characterized by strong synchronization, they are very prominent in power spectra of EEG signals. Although preictal states are not clearly visible in raw EEG signals, multiple studies confirmed the presence of distinct preictal states using spectral [132, 115, 133], as well as information measures. [134, 135, 136] For a detailed discussion, see [107] and [32].

Although power spectra capture relevant changes in frequency over time, they can be very noisy and contain outliers. We thus use nonnegative matrix factorization (NMF) [102, 101] to decompose power spectra into dominant time and frequency components, which are later used for seizure prediction.

To mitigate class imbalance, we employ synthetic minority over-sampling technique (SMOTE) [137], together with linear SVM with L1 regularization, to assign weights for contributions from each individual channel and eliminate uninformative channels. The method is applied to a part of the Freiburg EPILEPSIAE dataset [31], and compared to the Epilepsyecosystem dataset [114]. The developed method is computationally inexpensive and produces good results while providing insights into the structure of preictal states.

4.3 Materials and Methods[1]

4.3.1 Data preparation

Freiburg EPILEPSIAE dataset

The data consist of heterogeneous EEG recordings of five pre-surgical patients (one female; median age: 29.2) [Tab.4.1] and form a part of the bigger Freiburg EPILEPSIAE database.[31] Recordings are made at the University Medical Center Freiburg, over the course of several days (three to nine), between 2003 and 2009. The sampling frequency varies between 256Hz and 1024Hz. The electrodes that are used in the recordings include intracranial (depth, strip and grid) and surface electrodes, together with special electrodes (e.g. ECG, EMG and EOG), whose number varies between 31 and 122, depending on the diagnosis. In order to investigate preictal states thoroughly, only intracranial EEG recordings are used.

Since the ability to predict a seizure five minutes before its onset can be useful for patients with uncontrolled epilepsy [138], we focus on five minute intervals of preictal and interictal states. In the case of a preictal state, an interval of five minutes leading up to a seizure, with a 30 seconds seizure horizon is extracted. Seizure onsets are hand-labeled at the University Medical Center Freiburg. Since preictal states directly precede seizures, seizure prediction can be realized by classification between preictal and interictal states.

In the case of an interictal state five minutes intervals are extracted, which are at least 11 minutes before or after any other seizure. We refer to these intervals of extracted signals as individual measurement periods. The data are filtered with the Parks-McClellan optimal equiripple finite impulse response filter to remove 50Hz line noise.

The dataset is separated into training (70%) and validation set (30%) during a 100-fold cross-validation procedure.

Patient's number	age	sex	number of channels	sampling frequency (Hz)	number of preictal intervals	number of interictal intervals
1	34	male	48	256	16	88
2	37	female	26	512	6	44
3	18	male	94	1024	8	80
4	42	male	38	1024	6	110
5	15	male	91	256	14	9

Table 4.1. Detailed information about patients the from EPILEPSIAE database. [31] The number of preictal intervals is the same as the number of seizures.

Epilepsyecosystem dataset

The dataset consists of intracranial EEG recordings of three patients (all females; median age: 50). [Tab.4.2] Recordings are made at the St Vincent's Hospital in Melbourne, Australia as a part of the world-first clinical trial of the implantable

[1] A software implementation of the presented method is available online at: `https://github.com/ostojanovic/seizure_prediction`.

NeuroVista Seizure Advisory System. [110] In total, 16 electrodes are used for each patient and sampling frequency is 400Hz. The dataset consists of the public and the private (benchmark) set. Since labels of preictal and interictal states are known only for the public set, it is used for developing a model, while the benchmark set is used in the final stage for comparison with other algorithms for seizure prediction. [114]

Preictal intervals are ten minute segments which are cut out of recordings covering one hour prior to seizure with a five minute seizure horizon. (i.e. from 1:05 to 0:05 before seizure onset). Interictal intervals are also ten minute segments cut out from one hour of recording, which is at least four hours away from any seizure. Some of the files contain data dropouts which happen when the intracranial brain implant temporarily fails to record data. This manifests in zero values of iEEG across all channels at a given time sample. All files that contain more than 50% of data dropouts are excluded from the further analysis. For files that contain less than 50% of data dropouts, the corrupt data are deleted and the rest of the signal is concatenated. The data are filtered with the Butterworth infinite impulse response filter to remove 50Hz line noise.

The public dataset is separated into training (70%) and validation set (30%) during a 100-fold cross-validation procedure.

Patient's number	age	sex	number of preictal intervals	number of interictal intervals	number of files (benchmark set)	percentage of excluded files
1	22	female	225	500	162	14.9%
2	51	female	216	1688	941	7%
3	50	female	251	1896	679	1%

Table 4.2. Detailed information about the Epilesyecosystem dataset (after excluding corrupted files). [114] The number of preictal intervals is the same as the number of seizures in the public dataset, while for files in the benchmark dataset labels are not publicly known.

4.3.2 Deriving time and frequency components

To identify stereotypical behavior between and ahead of seizures, spectrograms of each channel [Fig.4.1] (for the Freiburg EPILEPSIAE dataset) are obtained using the multitaper method [139] with time windows of 10 seconds (which is calculated by using 50% overlap of a 20 seconds window). For the Epilepsyecosystem dataset, spectrograms of each channel are calculated using the Fast Fourier Transform. To correct for baseline activity across frequencies, relative power is calculated by dividing spectrograms of each channel by the average interictal spectrogram.

Due to the clinical setting and patients' diagnoses, the sampling frequency varies among different patients from the two datasets. As a result, the highest frequency in the spectrograms varies between 128Hz and 513Hz. However, this difference is unproblematic due to the fact that we develop patient-specific models. After obtaining spectrograms of every individual measurement period for every channel, they are visually inspected, and in the case of anomalies (e.g. electrode detachments, sudden amplitude jumps), excluded from the data.

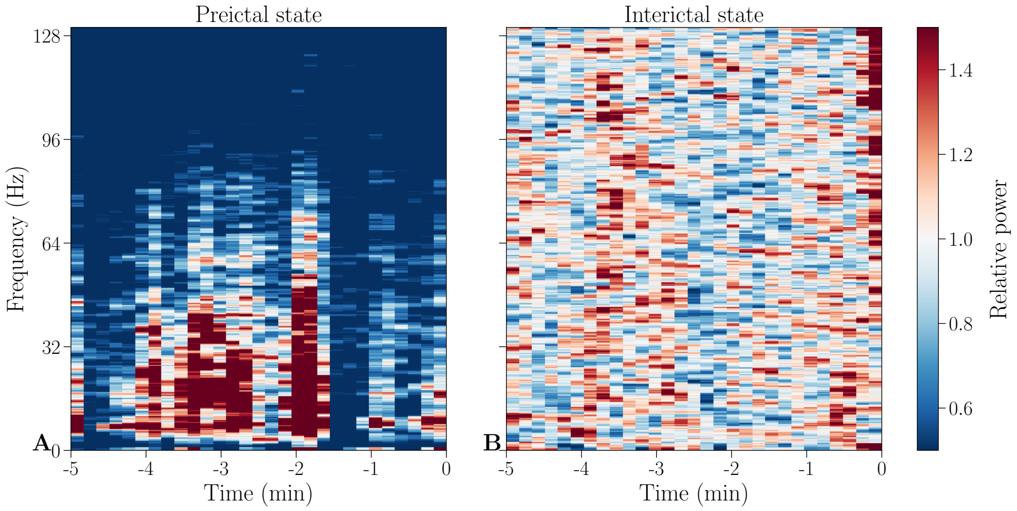


Figure 4.1. Example spectrograms of preictal and interictal states. Baseline corrected spectrograms of a preictal (**A**) and an interictal (**B**) individual measurement period of channel HR1 from patient 1. This channel and individual measurement period will be used throughout the paper for illustrative purposes, if not stated otherwise.

4.3.3 Time-frequency decomposition

To examine changes in power spectra, spectrograms of each channel and each individual measurement period are decomposed into a time and a frequency component using nonnegative matrix factorization. Originally proposed under the name “positive matrix factorization”, it is a variant of factor analysis [102], which is first used on environmental data [125] and later popularized in the application to face recognition under the current name. [101] For both tasks, NMF is successful in learning interpretable parts-based representation (e.g. concentrations of elements, as in [125] or parts of faces, as in [101]) and shown to perform better than independent component analysis, principal component analysis or vector quantization.[140, 141, 142] In the field of seizure prediction, NMF has been used to develop a method for automatic localization of epileptic spikes in children with infantile spasms [143] and for automatic detection and localization of interictal discharges.[144]

Nonnegative matrix factorization decomposes a nonnegative matrix V into two nonnegative low-rank matrices W and H [101]:

$$V \sim \tilde{V}_{n\times m} = W_{n\times r} \times H_{r\times m}$$

$$\tilde{V}_{ij} = \sum_{a=1}^{r} W_{ia} H_{aj}$$

The outer product $\tilde{V} = WH$ can be interpreted as a low rank parts-based approximation of the data in V.[101] We decide on a factorization of rank $r = 1$ to get the most constrained model with two vectors, one of which represents temporal evolution (time component H) and one of which represents distribution of frequencies (frequency component W). [Fig.4.2]

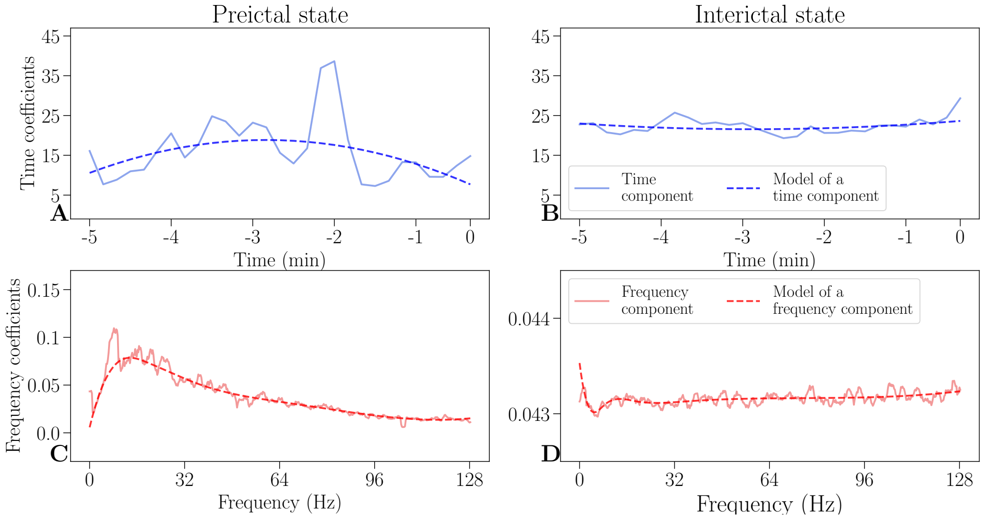


Figure 4.2. Time and frequency components and its models. An example of decomposed time (solid blue lines) and frequency components (solid red lines) and their respective models (dashed lines) of a preictal state (**A, C**), as well as an interictal state (**B, D**). In a preictal state, the time component (**A**) increases as a seizure is approaching, while the frequency component (**C**) has an increase in low frequencies. Both interictal components (**B, D**) are steady and are an order of magnitude lower than their respective preictal components (**A, C**).

To lessen the influence of outliers and to remove noise in the NMF components, they are modeled with smooth basis functions using robust regression. The time component is modeled by a polynomial of second order, while the frequency component is modeled by nonlinearly logarithmically spaced B-splines of sixth order to consider the frequency resolution which decreases in higher frequencies. [Fig.4.2] By modeling each component with smooth basis functions, the most relevant information is preserved in both domains, while noise is removed.

By calculating the outer product of modeled NMF components as shown in figure 4.3, time-frequency models can be reconstructed. They capture the most important information while leaving out the noise and thus provide simplified intermediate representation of the data, which can be visually compared to the corresponding spectrograms (see S1A Fig in the appendix). The coefficients of the modeled time and frequency components therefore convey relevant information about structure of both states.

4.3.4 Prediction and performance measures

To classify between preictal and interictal states, linear support vector machines [145] are used. We combine the coefficients of both of the modeled NMF components across all channels into a feature vector. For example, recordings of patient 1 in the EPILEPSIAE dataset contain 48 channels with 12 NMF

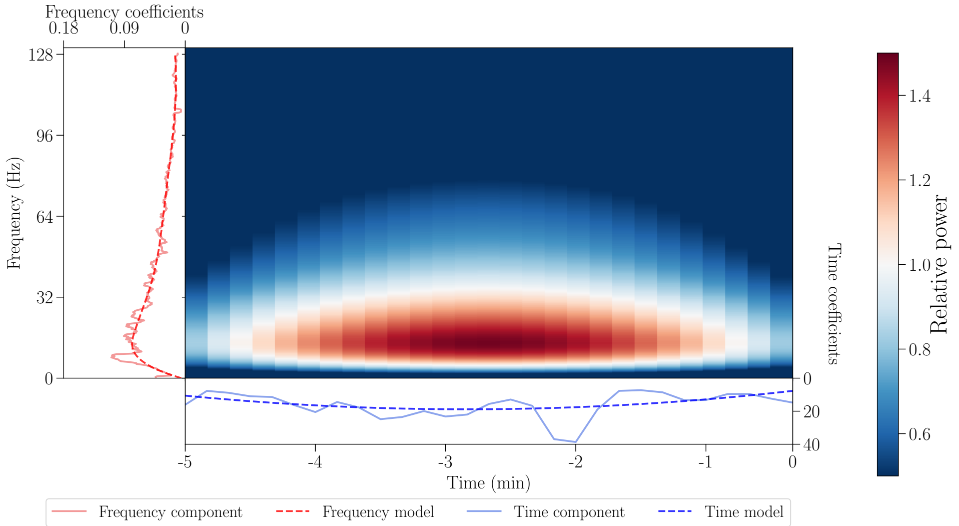


Figure 4.3. Obtaining a time-frequency model from the respective components. The NMF components are shown with solid red and blue lines for frequency and time, respectively, while their models are shown with dashed lines. The time-frequency model (center) is an outer product of modeled time and frequency components.

parameters (9 parameters for the frequency component and 3 parameters for the time component) each, leading to a dimensionality of $48 \cdot 12 = 576$. To account for the risk of overfitting due to the high number of features, L1 regularization is used. L1 regularization shrinks coefficients of less important features to zero by adding the absolute value of magnitude of coefficients as a penalty term to the loss function. [145]

In both datasets, interictal states are more frequent than the preictal ones, which leads to an imbalance of classes (c.f. Tab. 4.1 and 4.2). To account for this, the SMOTE oversampling technique is used. [137] It creates synthetic samples of the minority class, based on k neighboring points of minority samples (in our case $k = 5$). This means that the new synthetic preictal sample is created based on the five closest preictal samples.

To ensure good generalization of the algorithm, 100-fold cross-validation is used on a training set (70%) and a validation set (30%). Average measures (accuracy, sensitivity, specificity, positive and negative predictive values) are reported. Since the classifier should neither miss nor falsely predict a seizure, we report sensitivity sensitivity and specificity, as well as positive and negative predictive values. [146] In the benchmark dataset the area under the curve (AUC) is used for comparison among other algorithms.

Sensitivity is the probability of a positive test result among those having the target condition (i.e. the proportion of correctly classified preictal states), while specificity is the probability of a negative test result among those without the target condition (i.e. the proportion of correctly classified interictal states). [146] The positive predictive value (PPV) is the probability of the target condition, given a positive test result (i.e. the measure of how likely it is that, if the classifier predicts a preictal state, a patient is experiencing it), while the negative predictive value (NPV) is the probability of not having the target condition, given a negative test result (i.e. the measure of how likely it is that, if our classifier does not predict a preictal state, a patient is not experiencing it). [146] Full expressions are given below:

$$\begin{aligned}
\text{Accuracy} &= \frac{TP + TN}{\text{all samples}} \\
\text{Sensitivity} &= \frac{TP}{TP + FN} \\
\text{Specificity} &= \frac{TN}{TN + FP} \\
\text{PPV} &= \frac{TP}{TP + FP} \\
\text{NPV} &= \frac{TN}{TN + FN}
\end{aligned}$$

where:

TP is a number of samples classified as true positive

TN is a number of samples classified as true negative

FP is a number of samples classified as false positive

FN is a number of samples classified as false negative.

4.4 Results and discussion

4.4.1 Interpretability of the model

Figure 4.2 shows representative preictal and interictal components (of the EPILEPSIAE dataset), where the modeled NMF components show differences between the states. Model of the frequency component of a preictal state exhibits a peak of high activity in lower frequencies, relative to baseline activity. This is in line with previous findings of a structure below 30Hz (gamma range), which is informative for seizure prediction.[132, 133] These structural differences are also visible in recovered time-frequency models (see S2A Fig and S3A Fig in the appendix).

Average preictal and interictal components of all measurements and electrodes differ in both datasets, as shown in S4A Fig and S5A Fig in the appendix. On average, time components of preictal states in the EPILEPSIAE dataset have higher intensity, and frequency components show increase in lower frequencies (S4A Fig). Equivalent average components in the public Epilepsyecosystem show slightly different behavior. Time components of interictal states have somewhat higher intensity, and frequency components have an increase in lower as well as in higher frequencies. Since labels for the private Epilepsyecosystem dataset are not available, it is not possible to analyze the benchmark dataset in the same way.

Figure 4.4 shows normalized histograms of maximum values of frequency components of preictal and interictal states for both datasets. In the EPILEPSIAE dataset most preictal components have maximum in lower frequencies, and interictal states have maximum in both lower and higher frequencies (above 100Hz). On the other hand, most maxima of preictal and interictal components in the public Epilepsyecosystem dataset are below 50Hz as well as between 150Hz and 200Hz.

This difference in components between datasets can exist due to various reasons. The part of the EPILEPSIAE dataset used here might have too few measurements from an each patient. The Epilepsyecosystem dataset has more measurements, but it still contains data for only three patients. For a better assessment more data from different patients should be analyzed. In addition, it should be noted that the part of the EPILEPSIAE dataset used here contains data of pre-surgical patients and seizures recorded in this setting might not always be representative of typical epileptic seizures. As it is shown in [147], features of intracranial EEG signals show high variability after implantation of electrodes and spatial variability of lower frequency power bands across channels decreases over time. On the other hand, the Epilepsyecosystem dataset contains recordings from the world-first clinical trial of the human-implanted NeuroVista seizure advisory system [110], which might also be more distinguished than other clinical trials. Lastly, in the EPILEPSIAE dataset the 11-minutes buffer for interictal periods is used, which might be too short. The study in [148] reveals existence of “pre-cursors” to seizures (energy bursts in iEEG signals), which suggests that epileptic seizures might start hours in advance (also shown in [110]). Considering all of this, the best assessment of differences in preictal and interictal states would be in a closed-loop seizure prediction setting in real-time, for which the proposed method would, with appropriate adjustments (e.g. calculating spectrograms of consecutive time windows instead of short segments) be suitable.

4.4.2 Predictive performance

On the EPILEPSIAE dataset, similar accuracy is achieved for all patients (above 90%). The lowest performance is for the patient 5 (90.4%) and the highest for the patient 4 (100%), as shown in figure 4.5 and table 4.3. Sensitivity is between 0.8 and 1, while specificity ranges from 0.98 to 1, as can be seen in figure 4.5. A combination of high values of sensitivity and specificity is achieved for all patients. Similarly, positive predictive values are between 0.98 and 1, while negative predictive values are between 0.85 and 1 (c.f. figure 4.5 and table 4.3).

Predictions on the public Epilepsyecosystem dataset are lower than on the EPILEPSIAE dataset (around 70% for all patients; c.f. figure 4.5 and table 4.3). The lowest performance is for the patient 1 (74.1%) and the highest for the patient 3 (78.5%). Sensitivity, specificity, positive and negative predictive values for all patients are still higher than attainable results by a random classifier, but still considerably lower than on the EPILEPSIAE dataset, which can be seen in figure 4.5. Sensitivity is between 0.57 and 0.75, while specificity ranges from 0.73 to 0.82. Positive predictive values are between 0.63 and 0.81, and negative predictive values are between 0.75 and 0.77.

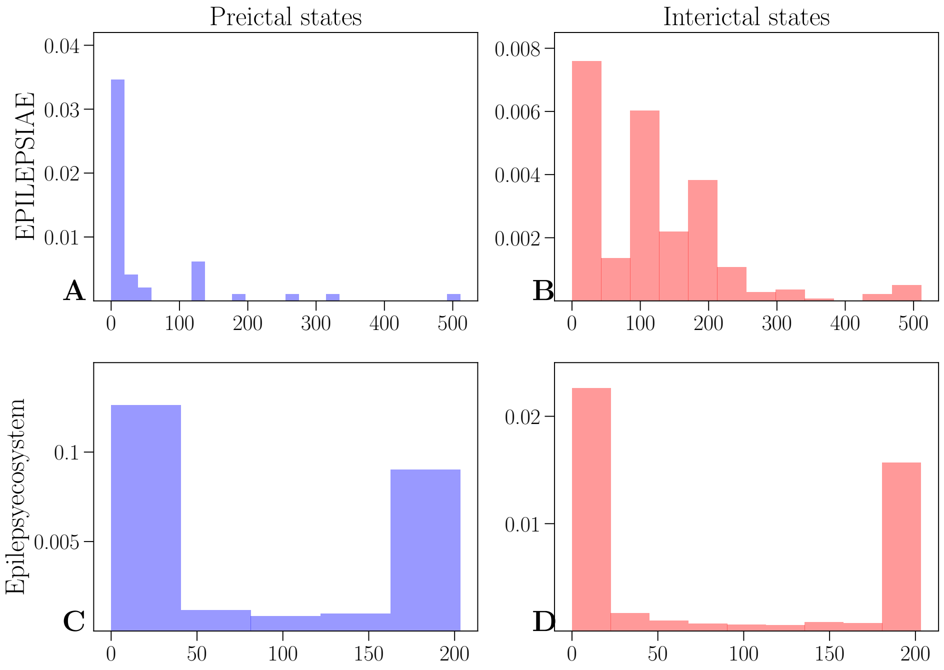


Figure 4.4. Distribution of maximum of frequency components. Results of the EPILEPSIAE dataset are shown in the upper row for preictal (**A**) and interictal states (**B**). The lower row shows results for the Epilepsyecosystem dataset (**C** for preictal and **D** for interictal states).

On the benchmark dataset, the highest achieved accuracy is for the patient 1 (71%), and the lowest for the patient 2 (61%). However, other performance measures drop significantly (sensitivity and positive predictive value are below 0.5). This drop in performance happens with most of other algorithms that are evaluated on the Epilepsyecosystem dataset [114], but the difference is not always as big. There might be various reasons for this. In general, it is the harder task to train a model on one dataset, and then evaluated it on the unseen set. Furthermore, the class imbalance between the sets might differ, which would explain the big difference between sensitivity and positive predictive value. It is also possible that SMOTE algorithm learns noise when oversampling the minority class in the public dataset. Finally, patients who have a higher seizure frequency (i.e. seizures per day) seem to have worse seizure prediction performance based on the original clinical trial. [110]

As mentioned in the *Prediction and performance measures*, the AUC is used for comparison with other algorithms on the benchmark set. The average reported AUC is 0.57 (0.62 for the patient 1, 0.52 for the patient 2 and 0.58 for the patient 3), which places the proposed algorithm on the 65th place (out of current 102 evaluated algorithms). For comparison, the algorithm with the best performance on the benchmark dataset (which is the combination of extreme gradient boosting, k-nearest neighbours, generalized linear model and linear SVM) has AUC of 0.8. [114]

The reasons for the overall lower performance on both Epilepsyecosystem datasets can lie in the fact that there are more seizures and more data per patient, making prediction possibly more challenging by potentially adding more variability to the data. It should also be noted that the data of three patients from the Epilepsyecosystem dataset correspond the ones whose seizures are the most difficult to predict [110].

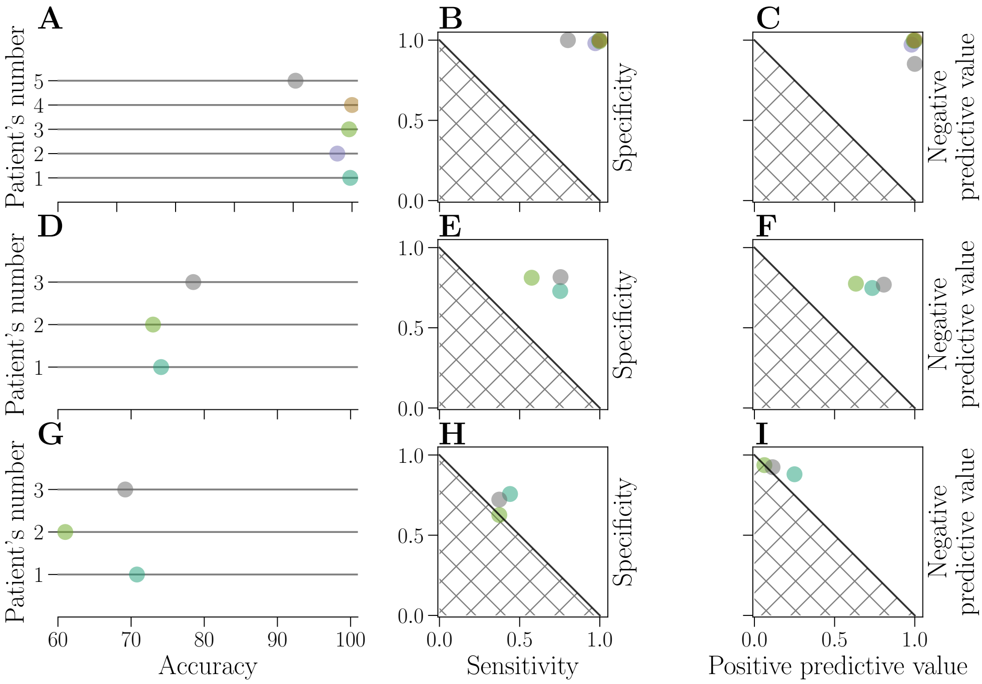


Figure 4.5. Evaluation of prediction performance. Results on the EPILEPSIAE dataset are shown in the upper row(**A-C**). Results on the public Epilepsyecosystem are shown in the middle row (**D-F**) and the results on the private Epilepsyecosystem dataset (benchmark) are shown in the lower row (**G-I**). Performance of each patient is represented by a circle, for accuracy (**A,D,G**), specificity-sensitivity plot (**B,E,H**) and negative and positive predictive value (**C,F,I**). Identical colors are used to represent each patient across all nine subplots. The hatched area represents results attainable by a random classifier.

4.5 Conclusion

Since patients with uncontrolled epilepsy prefer to be advised a few minutes before a seizure onset [138], we decided to use intervals of five minutes, extracted from longer recordings of the EPILEPSIAE dataset. However, this method is easily extensible to longer periods of time, since the length of intervals has no effect on dimensionality of modeled time components, which is shown by comparing the proposed method on the Epilepsyecosystem dataset.

Data from additional patients as well as more data from the same patient could, if available, lead to a better generalization of the model. This however is a

Patient's number	accuracy (%)	sensitivity	specificity	positive predictive value	negative predictive value
1	99.7	0.99	1	1	0.99
2	97.5	0.97	0.98	0.98	0.97
3	99.5	1	0.99	0.99	1
4	100	1	1	1	1
5	90.4	0.8	1	1	0.85
1	74.1	0.75	0.73	0.73	0.75
2	73	0.57	0.81	0.63	0.77
3	78.5	0.75	0.82	0.81	0.77
1	71	0.44	0.76	0.25	0.88
2	61	0.37	0.63	0.06	0.94
3	69.2	0.37	0.72	0.11	0.92

Table 4.3. Performance measures for all patients from the EPILEPSIAE dataset (upper section), from the Epilepsyecosystem public dataset (middle section) and Epilepsyecosystem benchmark dataset (lower section)
.

challenge for patient-specific models in general, where data from a single patient should suffice, and a large number of labeled training examples is not available.

Overall, this study demonstrates the use of nonnegative matrix factorization of power spectra for a seizure prediction task. The proposed model is conceptually simple, interpretable and has shown good accuracy on two representative datasets and lower performance on the benchmark set where improvements in the direction of coping with class imbalance should be made. A similar approach could be used for similar tasks such as detection of sleep stages in EEG or the detection of irregularities in ECG.

Acknowledgments

We are very grateful for early discussions with the team of the Freiburg Epilepsy Center, especially Prof.Dr.med. Andreas Schulze-Bonhage and Dr.-Ing. Matthias Dümpelmann, to helpful input from our colleague Johannes Leugering and to our anonymous reviewers who helped us to improve the paper.

Chapter 5

Example 2: predicting infectious diseases

Surveillance, prevention, and control of infectious diseases are fundamental tasks of public health agencies. Predicting the case counts of infectious diseases can inform public policies and thus help stop the spread, keeping the population healthy and minimizing potential deaths. [149, 150, 151] Predicting infectious diseases is about learning how many new infection counts will happen in a given time frame in the future in different spatial locations. [149] Such models predict cases in two dimensions: temporal, for which we use time series of cases as an input, and spatial, for which we use the location of new cases. [152, 153] However, infectious diseases are not static, and after interaction between cases, new ones will occur. Because of this, we want to make a model which can learn spatio-temporal interactions between cases [154, 155] and use this information for predictions. Besides spatio-temporal interactions, incorporating prior knowledge about diseases, seasonality, and trends is a great addition to spatio-temporal models. [156, 157]

The challenge in predicting infectious diseases is to make a model from which domain experts can learn disease dynamics and predict the number of cases and locations over time (since the public must know whether they need to take precautious measures). Depending on whether the disease is known and historical data are available or not, forecasting long into the future can be unreliable. In such cases, we usually make predictions for the present or near future, known as *nowcasting.* [158, 159] It helps us assess the situation since reporting is always delayed for a couple of days, i.e., the time between the incubation, laboratory tests, confirmation, and reporting in the database or, in general, delays due to the weekend and national holidays. [159, 160, 161] Because such epidemiological predictions can be used to inform policy-making, as in the coronavirus pandemic, the quality of data directly impacts the quality of decisions. Communicating and explaining predictions is key for public support of the measures (see *Appendix B, Visualizing the spread of infectious diseases using public health data*).

5.1 Properties of spatio-temporal data

Working with epidemiological data poses several challenges. Consider, for example, the epidemiological data we used here: time series of infection counts in Germany aggregated by administrative areas that correspond to NUTS3 areas (*Nomenclature of Territorial Units for Statistics*). [162] Collecting such data depends on administrative divisions in a country and data protection laws. For example, European Union (EU) member states collect epidemiological data on the NUTS3 level, which are then reported to European Centre for Disease Prevention and Control (ECDC). [163] NUTS3 regions are defined by Eurostat as areas that contain between 150,000 and 800,000 people, which is a relatively low resolution. [164] In Germany, this corresponds to the administrative level of counties (German *Landkreise*). While Germany only collects data on the NUTS3 level, other countries may also collect data on a finer level. For example, Belgium and the Netherlands collect infection counts on the level of municipalities [165, 166], whose average size is 20,000 and 50,000 people, respectively. [167, 168] This directly affects the spatial resolution of predicted infection counts.

Further, administrative divisions can have unusual shapes influenced by historical boundaries and may contain enclaves and exclaves (see a German exclave in Switzerland in Fig. 5.1). But diseases spread with human movement across the boundaries, making it more challenging to model the spatial spread of infectious diseases. To model spatio-temporal interactions of diseases "on the ground", we need to go from the abstract level of neighboring regions down to their actual boundaries in physical space. However, due to political and socio-economical transformations, administrative boundaries are also subject to change, e.g., as the population in cities grows, it is more likely that one area will split into multiple, and vice versa.

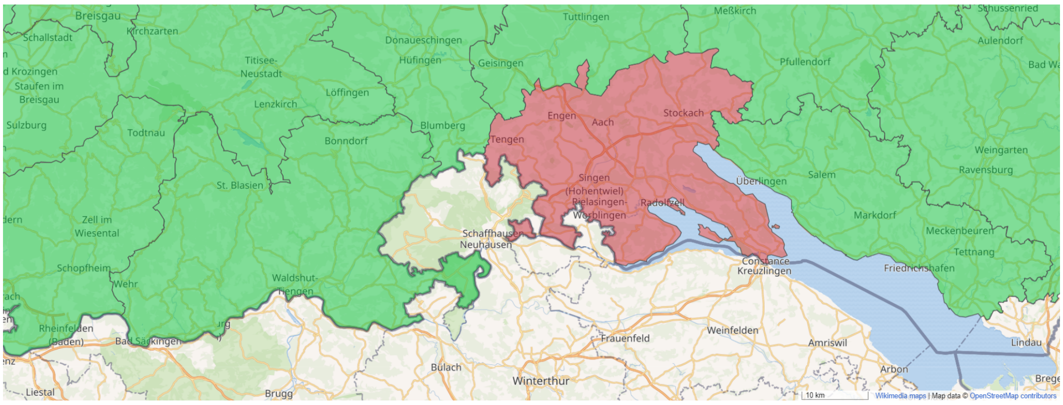


Figure 5.1. An example of an exclave. The German county Konstanz, marked in red contains an exclave Büsingen am Hochrhein, which is surrounded by Switzerland (German territory is colored in green). Image credits: OpenStreetMap under the Open Database Licence (ODbL) 1.0.

When making spatio-temporal predictions, we must first adjust historical data to changes in boundaries. Since NUTS3 doesn't change as often as some finer levels might, we did not have to adjust our data. If we had worked with a finer level in Germany, such as districts (*Gemeinde*), we would have to do so.

In addition to epidemiological data, demographics and socio-economical data are crucial for making predictions [169, 170, 171], as viruses can spread differently among age groups or people working different jobs. Data from social media can be a good and low latency addition to the model (e.g., the mentions of flu on Twitter). [172] However, one should exercise caution when using data from social media, since the demographics of users varies among platforms and can be significantly skewed. [173] Further, social media users show group dynamics that are not representative of usual human behavior, and it is generally unclear which data preprocessing and filtering has already been done internally at the companies. [174, 173].

5.2 Properties of public data

In this thesis, we used public data, which has some special properties. The Publications Office of the European Union defines public data as:

> *Open (Government) Data refers to the information collected, produced or paid for by the public bodies (also referred to as Public Sector Information) and made freely available for re-use for any purpose.* [175]

As such, public data depends on the laws and governments of the country of origin. Governments can decide the level at which data is collected, which directly influences the modeling process. Further, governments have power over which data to collect and which not, which in unstable political and economic situations can mean that there might be an incentive to misrepresent the data. [21]

Data privacy and people's attitudes toward sharing their data shape the locality of data collection and what the public is interested in getting from predictions of infectious diseases. A study in the UK showed that people are willing to share their data when the purpose is clear. [176, 177] In situations such as the coronavirus pandemic, communicating the current situation is crucial in ensuring the public's trust. [178] Although the public attitude towards collecting data varies among European countries [179, 180], publishing and communicating data on the level of NUTS3 seems to be generally accepted.[181] However, in Japan, people are more willing to share their data with private companies (like Apple) than with the government. [182] On the other hand, in South Korea, the public accepted maps that showed confirmed cases in the country, including personal information about the infected person and their movement. [183]

5.3 Modeling the spread of infectious diseases

In this project, we developed a general model of the spread and spatio-temporal interactions of infectious diseases. We tested the model on three diseases with different characteristics: campylobacteriosis, rotavirus, and Lyme borreliosis. We made one-week-ahead predictions for each county in Germany. In the model, we used spatio-temporal epidemiological data of case counts, with additional demographical and political information (e.g., whether counties belonged to eastern or western Germany).

Although domain knowledge is often used in statistical models to model the spread of infectious diseases [149, 184], they often don't show how parts of the model affect predictions. Lately, more machine learning methods combine epidemiological and other data types, such as environmental data, Google or Twitter trends, etc., to predict new cases. [185, 186, 187, 188] Interpretability of such methods usually depends on the type of prediction algorithm.

We designed an interpretable model to predict weekly case counts of three infectious diseases. Since we explicitly wanted to design one spatio-temporal model that could be adjusted for different diseases, we opted for a probabilistic generalized linear model with a Bayesian approach. Generalized linear models are suitable for predicting infectious diseases since they allow for non-Gaussian assumptions about the model, which is common in epidemiology. The Bayesian approach allows us to include domain knowledge as prior modeling assumptions, which later helps interpret results.

For several reasons, probabilistic modeling and Bayesian statistics are good choices for designing a model for predicting infectious diseases. First, Bayesian inference expliclitly combines prior assumptions with available data, which helps incorporate domain knowledge into the model. [189] The posteriors can be updated during learning, ensuring that models learn underlying relationships between variables. Because of this, Bayesian modeling can be used to develop interpretable machine learning models. [69] Second, Bayesian modeling can handle limited availability of data, as is often the case in epidemiology. Third, Bayesian probabilistic models give not only point predictions but rather probability distribution over predictions and the parameters of the model, which is useful for estimating the uncertainty of predictions. [189] This is important when basing high stakes decisions on these predictions and helps us decide whether we need to collect more data to make predictions less uncertain. Finally, Bayesian modeling allows to explicitly encode relations between variables. This leads to better descriptions of the model, especially if we want to learn more about the dynamics of the disease. [73, 80, 69] Since analytical solutions are complex and require a lot of computation power, sampling methods like Markov Chain Monte Carlo (MCMC) can be used, e.g., as implemented in the probabilistic programming software package PyMC3. [190]

Our model consists of basis functions that capture different factors of the spread of infections. We use spatio-temporal basis functions to learn interaction within geographical regions and over time. We also model the trends and seasonality of each disease and use demographic and region-specific information in the model. We employed MCMC sampling to learn Bayesian posterior distributions of model parameters and predcitions.

The model generalizes well and learns distrinct and interpretable spatio-temporal interaction kernels for each disease. The inferred kernels give insight into the dynamics of each disease and show how parts of the model affect prediction, which makes the model globally interpretable on a modular level.

Chapter 6

A Bayesian Monte Carlo approach for predicting the spread of infectious diseases

Contributions

This chapter was first published as: *Stojanović O, Leugering J, Pipa G, Ghozzi S, Ullrich A (2019) A Bayesian Monte Carlo approach for predicting the spread of infectious diseases. PLOS ONE 14(12): e0225838.*
https://doi.org/10.1371/journal.pone.0225838

Olivera Stojanović
Contributed equally to this work with: Olivera Stojanovic, Johannes Leugering
Roles: Conceptualization, Methodology, Project administration, Software, Visualization, Writing - original draft
Affiliation: Department of Neuroinformatics, Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany

Johannes Leugering
Contributed equally to this work with: Olivera Stojanović, Johannes Leugering
Roles: Conceptualization, Methodology, Software, Visualization, Writing - original draft
Affiliation: Department of Neuroinformatics, Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany

Gordon Pipa
Roles: Conceptualization, Supervision
Affiliation: Department of Neuroinformatics, Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany

Stéphane Ghozzi
Roles: Data curation, Resources, Validation, Writing - review & editing
Affiliation: Department of Infectious Diseases, Robert Koch Institute, Berlin,

Germany

Alexander Ullrich
Roles: Data curation, Resources, Supervision, Validation, Writing - review & editing
Affiliation: Department of Infectious Diseases, Robert Koch Institute, Berlin, Germany

Data Availability

All data and code used in this work is available at the public code repository at: `https://github.com/ostojanovic/BSTIM`.

Funding

The author(s) received no specific funding for this work.

Competing Interests

The authors have declared that no competing interests exist.

6.1 Abstract

In this paper, a simple yet interpretable, probabilistic model is proposed for the prediction of reported case counts of infectious diseases. A spatio-temporal kernel is derived from training data to capture the typical interaction effects of reported infections across time and space, which provides insight into the dynamics of the spread of infectious diseases. Testing the model on a one-week-ahead prediction task for campylobacteriosis and rotavirus infections across Germany, as well as Lyme borreliosis across the federal state of Bavaria, shows that the proposed model performs on-par with the state-of-the-art *hhh4* model. However, it provides a full posterior distribution over parameters in addition to model predictions, which aides in the assessment of the model. The employed Bayesian Monte Carlo regression framework is easily extensible and allows for incorporating prior domain knowledge, which makes it suitable for use on limited, yet complex datasets as often encountered in epidemiology.

6.2 Introduction

Public-health agencies have the responsibility to *detect*, *prevent* and *control* infections in the population. In Germany, the Robert Koch Institute collects a wide range of factors, such as location, age, gender, pathogen, and further specifics, of laboratory confirmed cases for approximately 80 infectious diseases through a mandatory surveillance system [162]. Since 2015, an automated outbreak detection system, using an established aberration detection algorithm [191], has been set in place to help *detect* outbreaks [192, 193]. However, *prevention* and *control* require quantitative *prediction* instead of mere *detection* of anomalies and thus prove more challenging. For logistical, computational and privacy reasons, epidemiological data is typically reported or provided in bulk, often grouped by calendar weeks and counties. Predictions thus have to be made about the number of cases per time-interval and region, based on a history of such measurements.

Since outbreaks can extend over multiple counties, states or even nations, spatio-temporal models are typically employed. Some approaches use scan statistics to identify anomalous spatial or spatio-temporal clusters [194, 195], while others model and predict case counts as time series or point processes [196, 197]. A major advantage of such predictive models is the additional insight they can provide into the factors contributing to the spread of infectious diseases.

In general, we distinguish four qualitatively different classes of predictive features: *spatial*, *temporal*, *spatio-temporal* and *(spatio-temporal) interaction* effects. The former three are purely functions of space, time or both, modeling *seasonal* fluctuations and *trends*, *geographical* influences or localized time-varying effects, such as *region-specific demographics* or *legislation*, respectively. The latter is an autoregressive variable that captures how an observed infection influences the number of further infections in its neighborhood over time, which depends on differences in *patients' behavior*, *transmission vectors*, *incubation times* and *duration* of the respective diseases. Even in the absence of direct contagion, previously reported cases can provide valuable *indirect* information for predicting future cases through latent variables. The effect on the expected number of cases at a given place and time due to interactions can thus be expressed as a (unknown) function of spatial and temporal distance to previously reported cases. Particularly for regions with less available historic data or those strongly influenced by their neighbors, e.g. smaller counties close to larger cities [198], incorporating the county's and its neighbors' recent history of case counts can improve predictions.

The state-of-the-art spatio-temporal *hhh4* method [196, 199] assumes aggregated case counts to follow a Poisson or Negative Binomial distribution around a mean value determined by "epidemic" and "endemic" components. The epidemic component can capture the influence of previous cases from the same or neighboring counties, e.g. potentially weighted by the counties' adjacency order, while the endemic component models the expected baseline rate of cases.

For *not aggregated* data, the more general *twinstim* method [196] models the interaction effects due to individual cases by a self-exciting point process with predefined continuous spatio-temporal kernel, rather than through a binary neighborhood relation as in the *hhh4* model. Optimizing such a kernel for a specific dataset provides an opportunity to incorporate or even infer information about the infectious spread of the disease at hand. Using such smooth spatial kernel functions in favor of e.g. neighborhood graphs between geographical regions has the additional benefit, that it can also be applied in domains where the shape and neighborhood relation between such regions is complex. For example within Germany counties can contain enclaves, e.g. cities that represent a county of their own, or even be composed of disjoint parts.

In the following, we present a Bayesian spatio-temporal interaction model (referred to as BSTIM), as a synthesis of both approaches: a probabilistic generalized linear model (GLM) [200] predicts aggregated case counts within spatial regions (counties) and time intervals (calendar weeks) using a history of reported cases, temporal features (seasonality and trend) and region-specific as well as demographic information. Like for the *twinstim* method, interaction effects are modeled by a continuous spatio-temporal kernel, albeit parameterized with parameters inferred from data. Since the aggregated reporting of case counts per calendar week and county leaves residual uncertainty about the precise time and location of an individual case, we model times within the respective week and locations within the respective county as latent random variables. Monte Carlo methods are employed to evaluate posterior distributions of parameters as well as predictions, which are subsequently used to assess the quality of the model.

For three different infectious diseases, *campylobacteriosis*, *rotaviral enteritis* and *Lyme borreliosis*, the interpretability of the inferred components, specifically the interaction effect kernel, is discussed and the predictive performance is evaluated and compared to the *hhh4* method.

6.3 Materials and methods

We evaluate both the proposed BSTIM as well as the *hhh4* reference model on a one-week-ahead prediction task, where the number of cases in each county is to be predicted for a specific week, given the previous history of cases in the respective as well as surrounding counties. Instead of point estimates, we are interested in a full posterior probability distribution over possible case counts for each county and calendar week – capturing both aleatoric uncertainty due to the stochastic nature of epidemic diseases as well as epistemic uncertainty due to limited available training data. The data for this study is provided by the Robert Koch Institute, and consists of weekly reports of case counts for three diseases, campylobacteriosis, rotavirus infections and Lyme borreliosis. They are aggregated by county[1] and collected over a time period spanning from the 1st of January 2011 (2013 for borreliosis) to the 31st of December 2017 via the *SurvNet* surveillance system [162]. Aggregated case counts of diseases with mandatory reporting in Germany can be downloaded from https://survstat.rki.de. For each of the three diseases, the data preceding 2016 is used for training the model, while the remaining two years are used for testing. A software implementation of the BSTI Model presented here is available online at https://github.com/ostojanovic/BSTIM.

6.3.1 The BSTI Model

The proposed model is optimized to predict the number of reported cases in the future (e.g. the next week), based on prior case counts. Since epidemiological count data is often overdispersed relative to a Poisson distribution [201], i.e. the variance exceeds the mean, we assume counts are distributed as a Negative Binomial random variable around an expected value $\mu(t, x)$ that varies with time

[1] We use the term "county" to generally refer to rural counties (*Landkreise*) and cities (*kreisfreie Städte*) as well as the twelve districts of Berlin (*Bezirke*).

(t) and space (x), and with a scale parameter[2] $\alpha \geq 0$. The Negative Binomial distribution has been successfully used in epidemiology [201, 202, 203], since its variance $\mathbb{V} = \mu + \alpha\mu^2$ allows to model overdispersion in the data for $\alpha > 0$, while including the Poisson distribution as a special case for $\alpha \to 0$.

We further assume that the relationship between each feature $f_i(t,x)$ and the expected value $\mu(t,x)$ can be expressed in a generalized linear model of the Negative Binomial random variable $Y(t,x)$ using the canonical logarithmic link function. A half-Cauchy distribution is used as a weakly informative prior [204] to enforce positivity of the dispersion parameter of the residual Negative Binomial distribution. For all other parameters, Gaussian priors with zero mean and standard deviation 10 are chosen. Since the linear predictor of the generalized linear model combines qualitatively different types of data, specifically interaction effects and exogenous features such as temporal or demographical information, we employ sensitivity analysis to verify that the chosen (relative) scales for the priors do not unduly influence the inferred parameters. To this end, we systematically vary the standard deviation of the prior distribution for the interaction effect coefficients over the values 0.625, 2.5, 10, 40 and 160. Since we only observe negligible changes in the posterior parameter distributions (see supplementary S4C Fig through S6C Fig) and resulting predictions (not shown here) for standard deviations 10 and above, we conclude that the chosen Normal distribution with standard deviation 10 constitutes an adequate weekly informative prior. The full probabilistic model for training can thus be summarized as follows:

$$\alpha \sim \text{HalfCauchy}(\gamma = 2) \tag{6.1}$$

$$W_i \sim \text{Normal}(\mu = 0, \sigma = 10) \tag{6.2}$$

$$\mu(t,x) = \exp\left(\sum_{i=1}^{N} W_i f_i(t,x)\right) \cdot \epsilon(t,x) \tag{6.3}$$

$$Y(t,x) \sim \text{NegBin}(\mu(t,x), \alpha) \tag{6.4}$$

where:

α is a dispersion parameter

N is the total number of used features

W_i are model weights

$f_i(t,x)$ are features varying in time and space

$\epsilon(t,x)$ is the exposure varying in time and space

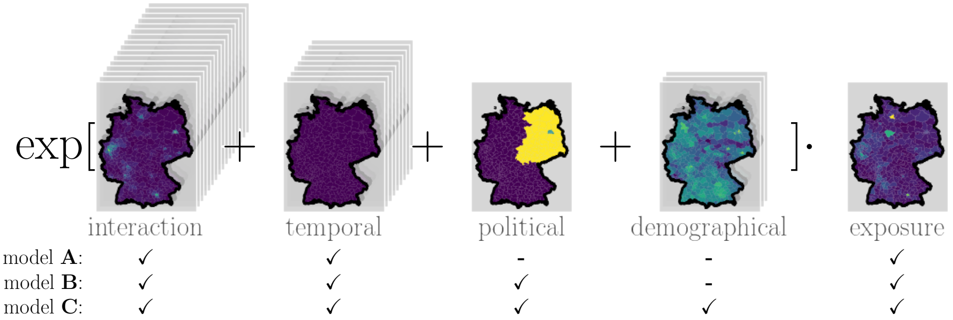
t refers to a time-interval (i.e. one calendar week)

[2] Due to its common use in combinatorics, the Negative Binomial distribution is often formalized in terms of parameters r, representing the number of failures in a hypothetical repeated coin flip experiment, and p, representing the success probability in each trial. This can be trivially extended to real valued coefficients, and reparameterized in terms of μ and α by setting $\mu \to {}^{pr}/_{1-p}$ and $\alpha \to {}^{1}/_{r}$.

x refers to a spatial region (i.e. one county)

For prediction, the priors over the dispersion parameter and weights are replaced by the corresponding posterior distribution inferred on the training set.

A schema of our model is shown in Fig.6.1. To capture the interaction effects between different places over time, a continuous spatio-temporal kernel is estimated through a linear combination of 16 basis kernels. The individual contribution due to each of these basis kernels is included into the model as a feature. Four temporal periodic *basis functions* are used to capture seasonality and five sigmoid *basis functions* (one for each year of available training data) to capture temporal trends. Four region-specific features (ratio of population in a county belonging to three age groups and one political component) are used, which results in 29 features. In addition, the logarithm of the population of each county in the respective year is used as a scaling parameter (exposure) ϵ.



	interaction	temporal	political	demographical	exposure
model **A**:	✓	✓	-	-	✓
model **B**:	✓	✓	✓	-	✓
model **C**:	✓	✓	✓	✓	✓

Figure 6.1. Model scheme. Exemplary contributions from different features, grouped into interaction, temporal, political and demographical components, each evaluated in all counties in Germany for campylobacteriosis in the week 30 of 2016. Each county's total population is always included as an exposure coefficient. We consider three models of increasing complexity, A, B and C, that differ in whether features are included (✓) or not (-). Information about the shape of counties within Germany is publicly provided by the German federal agency for cartography and geodesy (Bundesamt für Kartographie und Geodäsie) (GeoBasis-DE / BKG 2018) under the dl-de/by-2-0 license.

For example, given one parameter sample $w = [w_1, \dots, w_{29}]$, inferred from the training set of campylobacteriosis case counts, the conditional mean prediction within county x during week t is determined as follows:

$$\mu(t,x) = \exp\left(\underbrace{\sum_{i=1}^{16} w_i f_i(t,x)}_{\text{interaction}} + \underbrace{\sum_{i=17}^{20} w_i f_i(t)}_{\text{periodic}} + \underbrace{\sum_{i=21}^{25} w_i f_i(t)}_{\text{trend}} + \underbrace{\sum_{i=26}^{29} w_i f_i(t,x)}_{\text{region-specific}}\right) \cdot \underbrace{\epsilon(t,x)}_{\text{exposure}} \tag{6.5}$$

6.3.2 Monte Carlo sampling procedure

The model described above determines the posterior distribution over parameters by the data-dependent likelihood and the choice of priors. We want to capture this parameter distribution in a fully Bayesian manner, rather than summarize it by its moments (ie. mean, covariance, etc.) or other statistics. Since an analytic solution is intractable, we use Markov Chain Monte Carlo (MCMC) methods to generate unbiased samples of this posterior distribution. These samples can be used for evaluation of performance measures (here deviance and Dawid-Sebastiani score; cf. section *Predictive performance evaluation and model selection*), visualization or as input for a superordinate probabilistic model.

Our model combines features that can be directly observed (e.g. demographic information) with features that can only be estimated (e.g. interaction effects, due to uncertainty caused by data aggregation). To integrate the latter into the model, we generate samples from the distribution of interaction effects features as outlined in section *Interaction effects*.

The sampling procedure generates samples from the *prior* distribution over parameters and combines them with training data and our previously generated samples of the interaction effect features to produce samples of the *posterior* parameter distribution. These samples from the inferred joint distribution over *parameters* are then used to generate samples of the posterior distribution of model *predictions* for testing data.

We employ a Hamiltonian Monte Carlo method, No-U-Turn-Sampling[205], implemented in the probabilistic programming package *pyMC3*[190]. To evaluate proper convergence of the sampling distribution to the desired (but unknown) posterior distribution, four independent Markov chains are generated and their marginal distributions compared using the Gelman-Rubin diagnostic $\hat{R}$ [206], which assesses the relation between the within-chain and the between-chains variance.

6.3.3 Interaction effects

Each reported case provides valuable information about the expected number of cases to come in the near future and close proximity. We suppose that this effect of an individual reported infection on the rate of future (reported) infections in the direct neighborhood can be captured by some unknown function $\kappa(d_{\text{time}}(t_\star, t_k), d_{\text{geo}}(x_\star, x_k))$, which we refer to as *interaction effect kernel* in the following, where (t_k, x_k) refer to the time and location of the k-th reported case and $(t_\star, x_\star)$ represent the time and location of a hypothetical future case. Here, $d_{\text{geo}}(x, y)$ represents the geographical distance between two locations x and y, whereas $d_{\text{time}}(t, s)$ denotes the time difference between two time points t and s. Thus, $\kappa(\cdot, \cdot)$ is a radial, time- and location-invariant kernel, depending only on the spatial and temporal proximity of the two (hypothetical) cases. For the sake of simplicity, we assume that interaction effects due to individual infections add up linearly.

Since κ is not known a-priori for each disease, we wish to infer it from data. To this end, we approximate it by a linear combination of spatio-temporal basis kernels $\kappa_{i,j}$ with coefficients w_i that can be inferred from training data:

$$\begin{aligned}\kappa(\triangle t, \triangle x) \approx & \hat{\kappa}(\triangle t, \triangle x) := \sum_i w_i \kappa_{I_i,J_i}(\triangle t, \triangle x) \qquad (6.6)\\ \text{where} \quad & I_i := \lceil i/4 \rceil, \quad J_i := (i-1) \bmod 4 + 1\end{aligned}$$

As the basis functions for the interaction effect kernel, we choose the products $\kappa_{i,j}(\triangle t, \triangle x) := \kappa_i^T(\triangle t) \cdot \kappa_j^S(\triangle x)$ between one temporal (κ_i^T) and one spatial factor (κ_j^S), each (cf. Fig.6.2). As temporal factors, we use the third order B-spline basis functions $\kappa_i^T = N_{i,3}$ for $i = \{1, 2, 3, 4\}$ as defined in [207], with the knot vector $[0, 0, 7, 14, 21, 28, 35]$ (measured in days). The multiplicity 2 of the first knot enforces $\kappa_1^T(0) = 0$. This results in four smooth unimodal functions, spanning the overlapping time interval from zero to two weeks, zero to three weeks, one to four weeks and two to five weeks after a reported case, respectively. Outside these intervals, the functions are identically zero. Acausal effects (i.e. the influence of a reported case on hypothetical other cases reported at an earlier time) as well as effects more than five weeks after a reported case are thus excluded. This accounts for the typical incubation times for campylobacteriosis [208] and rotavirus infections [209], and early symptoms of Lyme Borreliosis [210], as well as potential reporting delays. As spatial factors, we use exponentiated quadratic kernels (i.e. univariate Gaussian functions) centered at a distance of 0km to a reported case, with shape parameters σ of 6.25km, 12.5km, 25.0km, and 50.0km. These spatial kernels are wide enough to cover the typical daily commuting distances within Germany, which amount to 25km or less for the majority of commuters [211], while being narrow enough to capture only local effects. See Fig.6.2 for an illustration of how the basis functions $\kappa_{i,j}$ are constructed.

Since the contributions of individual cases are assumed to sum up linearly, the total influence of all cases that were previously reported at times and places (t_k, x_k), $k \in 1 \ldots n$ onto the expected rate of cases reported at a later time t and location x is given by:

$$\begin{aligned}&\sum_{i=1}^{16} w_i f_i(t,x) \quad \text{where}\\ &f_i(t,x) := \sum_{k=1}^{n} \kappa_{I_i,J_i}(d_{\text{time}}(t,t_k), d_{\text{geo}}(x,x_k)) \qquad (6.7)\end{aligned}$$

Each $f_i(t,x)$ for $i \in \{1, \ldots, 16\}$ is a spatio-temporal function that depends on all cases reported prior to t, providing us with a total of 16 features for modeling interaction effects. By determining the corresponding coefficients w_i, the fitting procedure thus allows us to infer an interaction effect kernel $\hat{\kappa}$ in a 16-dimensional parameterized family from data. It should be noted, however, that since the basis functions $\kappa_{i,j}$ capture strongly correlated and possibly redundant information, the effective number of degrees of freedom may be well below 16. Since we work with aggregated data at a spatial resolution of counties and a temporal resolution of

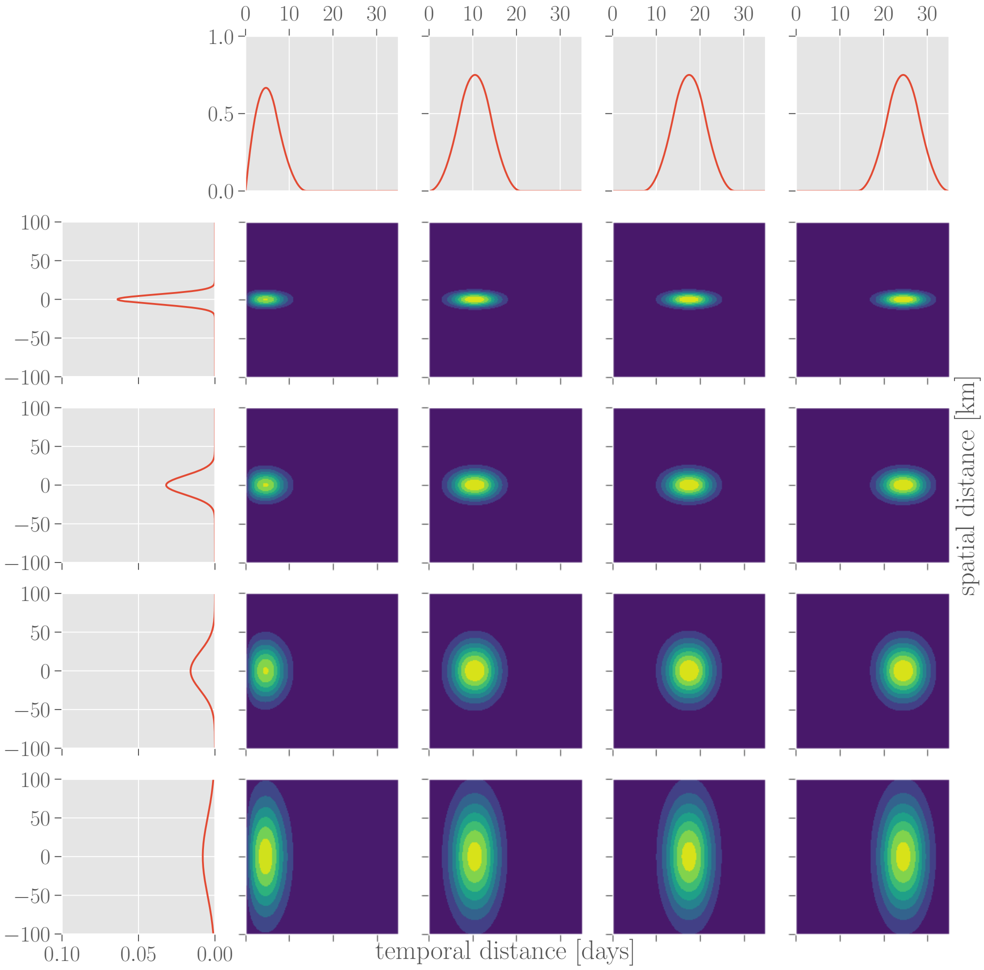


Figure 6.2. Spatial and temporal basis functions for interaction kernel. The inferred interaction kernel is composed of a linear combination of spatio-temporal basis functions (four-by-four grid of contour plots), each of which is a product of one spatial (left column) and one temporal factor (top row).

calendar weeks, the exact time and location of an individual case report, as well as time and location of a hypothetical future case, are conditionally independent random variables given the county and week in which they occur. Because of this epistemic uncertainty, the features $f_i(t, x)$ derived in equation 6.7 are thus random variables themselves. To deal with this uncertainty, the *twinstim* model proposed in [196] suggests to replace these features by their expected values, which can be numerically approximated efficiently. Here, instead of using such point-estimates, which might lead the model to underestimate its uncertainty, we want to incorporate the features $f_i(t, x)$ directly into our probabilistic model and thus need to account for their full probability distribution.

While this distribution is intractable to calculate analytically, we can generate unbiased samples from it through rejection sampling: For a case reported in a given calendar week and county, possible sample points of a precise time and location can be independently generated by uniformly drawing times from within the corresponding week and locations from a rectangle containing the county, rejecting points that fall outside the county's boundary. By randomly drawing a sample time and location for each reported case, we can thus generate an unbiased sample of the (unavailable) data prior to aggregation that accurately reflects the uncertainty caused by the aggregation procedure. Using these resulting sample times and locations in place of t_k and x_k in equation 6.7 yields unbiased samples of the features $f_i(t, x)$, which are in turn used when generating samples of the model's posterior parameter distribution (cf. section *Monte Carlo sampling procedure*).

It bears repeating that what we refer to as interaction effect features in this paper are thus in fact latent random variables due to the epistemic uncertainty caused by aggregated reporting of infections by counties and calendar weeks.

6.3.4 Additional features

Infection rates vary in time due to natural processes, such as seasons and climate trends, evolution of pathogens and immunization of the population, as well as societal developments such as scientific and technological advancement and medical education. Within Germany these effects may not differ much across space and can thus be included into the model as feature functions $f_i(t)$ that only depend on time. For modeling yearly seasonality, four sinusoidal basis functions (ie. $\sin(2\pi \cdot t \cdot \omega_{\text{yearly}})$, $\sin(4\pi \cdot t \cdot \omega_{\text{yearly}})$, $\cos(2\pi \cdot t \cdot \omega_{\text{yearly}})$, $\cos(4\pi \cdot t \cdot \omega_{\text{yearly}})$) are used as temporal periodic components, where $\omega_{\text{yearly}} = (1 \text{ year})^{-1}$. Slower time-varying effects are subsumed in a general trend modeled by a linear combination of one logistic function $\big($ie. $\big(1 + \exp\big(-\frac{t-\tau_i}{2} \cdot \omega_{\text{weekly}}\big)\big)^{-1}\big)$ centered at the beginning of each year (τ_i) with slope $1/2\ \omega_{\text{weekly}}$, where $\omega_{\text{weekly}} = (1 \text{ week})^{-1}$.

Due to the historical division between eastern and western Germany, and their different developments, some structural differences remain, such as unemployment rate, density of hospitals and doctors, population density, age structure etc. [212, 213] To account for such systematic differences, a political component, which we refer to as the *east/west component* in the following, is introduced which labels all counties that were part of the former German Democratic Republic as 1 and counties that were part of the Federal Republic of Germany as 0. Since Berlin itself was split into two parts, yet todays counties don't accurately reflect this historic division, counties within Berlin are labeled with an intermediate value of 0.5.

Since diseases can affect children and elderly in different ways, yearly demographic information about each county is incorporated into the model. The logarithm of the fraction of population belonging to three age groups (ages $[0-5)$, $[5-20)$ and $[20-65)$) is used. The total population of each county acts as a scaling factor for the predicted number of infections[3].

[3] The age group of 65 years and above accounts for the remaining share of the population and thus is a redundant variable with respect to the other three age groups and the total population.

6.3.5 Predictive performance evaluation and model selection

To evaluate the predictive performance of the model, forecasts of the number of infections are made one calendar week ahead of time for each disease and each county. To determine the relevance of different features, model selection is performed on the training dataset between three models of different complexity [Fig.6.1]:

model A - includes interaction and temporal (periodic and trend) components,

model B - includes interaction, temporal and political components,

model C - includes interaction, temporal, political and demographic components.

The Widely Applicable Information Criterion [214](WAIC, also referred to as *Watanabe-Akaike information criterion*, is applied to the posterior distribution over parameters and predictions from the training set to determine which combination of features (i.e. model A, B or C) minimizes the generalization error. Similar to the deviance information criterion, WAIC assesses the model's ability to generalize by estimating the out-of-sample expectation, while penalizing a large *effective* number of parameters. This is relevant here since modeling interaction effects introduces multiple features that capture redundant information. However, rather than evaluating the log-posterior at a parameter point-estimate, the WAIC calculates the empirical mean over the entire posterior distribution, which leads to a better estimate of the out-of-sample expectation [215], and is therefore ideally suited for sampling-based approaches.

Different error measures are applied to evaluate the fit of the predictive distribution for the test set to observations. Deviance of the Negative Binomial distribution (i.e. the expected difference between the log-likelihood of observations and the log-likelihood of the predicted means) is used as a likelihood-based measure and the Dawid-Sebastiani score (a covariance-corrected variant of squared error, cf. [216]) is included as a distribution-agnostic proper scoring rule.

To evaluate the performance of the model presented here as well as an *hhh4* model implementation for reference, we compare the resulting distributions of scores across counties.

6.3.6 The *hhh4* model reference implementation

We use an *hhh4* model for Negative Binomial random variables, implemented in the R package "surveillance"[217], with a mean prediction composed of an epidemic and an endemic component. The epidemic component is a combination of an autoregressive effect (models reproduction of the disease within a certain region) and a neighborhood effect (models transmission from other regions). The endemic component models a baseline rate of cases due to the same features as described above. The reference model is trained and evaluated on the same datasets as the BSTIM.

6.4 Results and discussion

Testing models of varying complexity (see Fig.6.1) reveals that the most complex model (model complexity C, including interaction effects, temporal, political as well as demographical features) generalizes best as measured by WAIC for all three different tested diseases (campylobacteriosis, rotavirus and borreliosis). [Tab.6.1] For the remainder of this text, we thus focus only on the full model variety C. The posterior parameter distribution inferred from the training data can be analyzed in itself, which provides valuable information about the disease at hand as well as the suitability of the model. Subsequently, it is used to generate one-week-ahead predictions for the test data.

model	campylobacteriosis	rotavirus	borreliosis
A	423279.3	349182.37	31359.62
B	420172.1	339143.27	(31359.62)
C	420010.64	338219.46	30643.49

Table 6.1. Training set WAIC scores for the three tested diseases and the three levels of model complexity. Since for borreliosis the model is trained and evaluated only within the western state of Bavaria, the east/west feature is constantly zero, and the models A and B thus coincide.

For each model configuration and disease, the sampling procedure is run until a total of 1000 valid samples of the joint posterior distribution have been generated, which each requires approximately four hours of run-time on a conventional desktop machine[4]. The sampling procedure converges to the same posterior for all independent chains, as can be seen by inspecting the posterior marginal distributions of each parameter in the supplementary figures S1C Fig to S3C Fig, which is quantified by the Gelman-Rubin diagnostics shown in supplementary figure S10C Fig.

6.4.1 The inferred model

The procedure outlined above produces samples from the posterior parameter distribution, which in turn provides a probability distribution over interaction kernels. Due to the large number of free parameters (16) involved (see Fig.6.2), the family of parameterized kernels is flexible enough to capture different disease-specific interactions in time and space. Despite the fact that much more complex interaction effect kernels could be learned, the kernels inferred from data appear to factorize into a specific spatial and temporal profile for each disease. The mean interaction kernel for campylobacteriosis (see Fig.6.3, 1A) shows the furthest spatial influence over up to 75 km, whereas rotavirus (see Fig.6.3, 2A) and borreliosis (see Fig.6.3, 3A) are more localized within a radius of up to 25 km. Borreliosis exhibits longer lasting interaction effects, extending up to four weeks. Despite the fact that borreliosis is not contagious between humans, this is consistent with a pseudointeraction effect due to a localized, slowly changing latent

[4] Utilizing 4 cores of an AMD Ryzen 5 1500x processor.

variable such as the prevalence of infected ticks or other seasonal factors. The kernel for campylobacteriosis shows a clear drop in the third week after an infection, which might indicate recovery from the disease, but we advise caution against overinterpretion of this negative interaction.

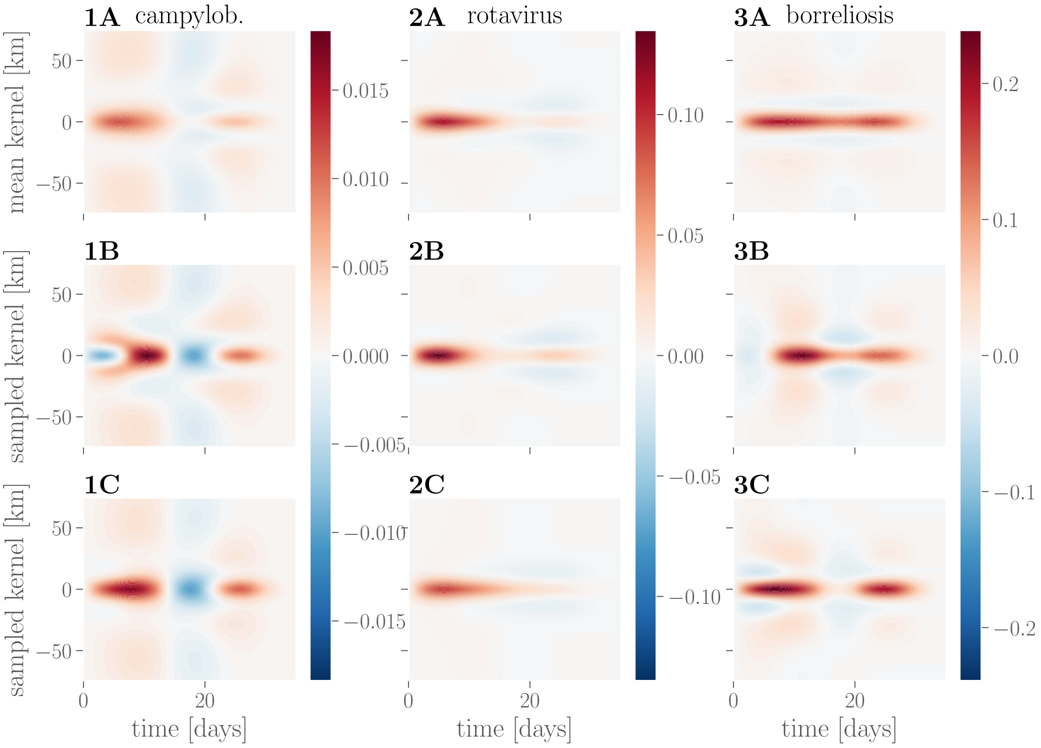


Figure 6.3. Learned interaction effect kernels. Kernels for campylobacteriosis are shown in **1A-C**, for rotavirus in **2A-C** and for borreliosis in **3A-C**. Mean interaction kernels are shown in the **row A**, while **rows B and C** show two random samples from the inferred posterior distribution over interaction kernels.

Looking at individual samples from the respective kernel distributions (see Fig.6.3, rows B and C) reveals a degree of uncertainty over the precise kernel shape for the different diseases: while there is little variation in the kernel shape inferred for rotavirus, there is uncertainty about the temporal profile of interactions for campylobacteriosis.

The seasonal components (see Fig 6.4) for campylobacteriosis and borreliosis show a yearly peak in July and June, respectively. In the case of rotavirus the incidence rate is higher in spring with a peak from March to April. The learned trend components capture the disease-specific baseline rate of infections, which remains stable throughout the years 2013 to 2016. While there is little uncertainty in the seasonal component, there is a high degree of uncertainty in the constant offset of the trend component. The effect of combining both contributions within the model's exponential nonlinearity results in higher uncertainty around larger values.

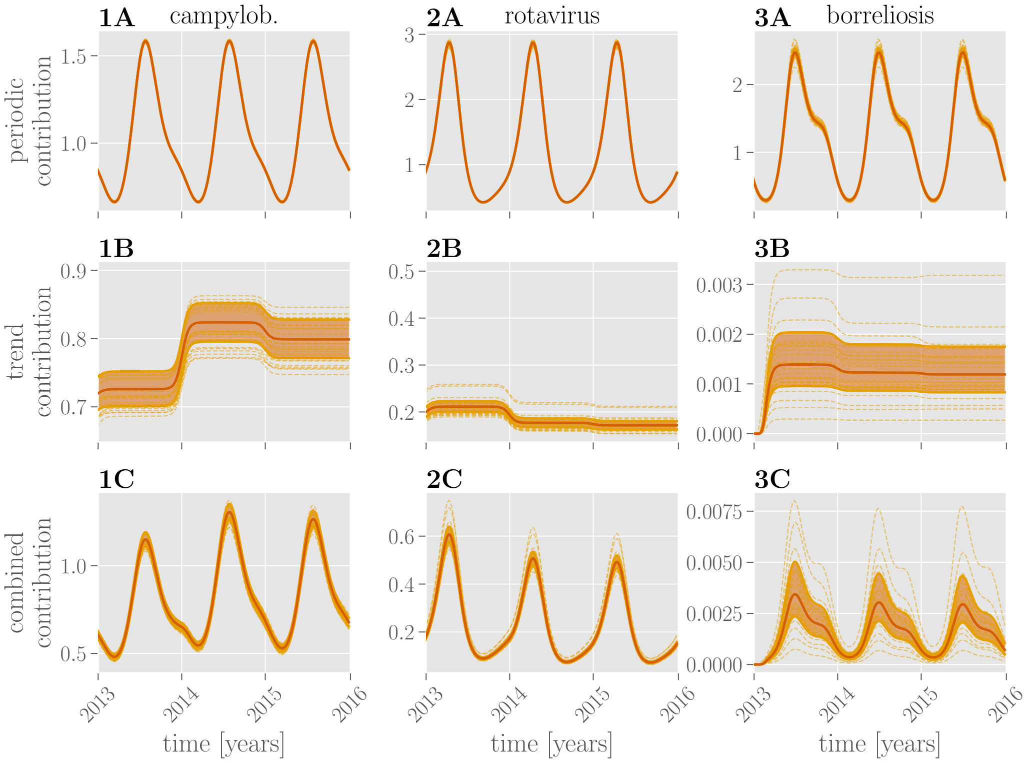


Figure 6.4. Learned temporal contributions. Periodic contributions over the course of three years (2013-2016) for all three diseases are shown in the **row A**, trend contributions in the **row B** and their combination in the **row C**. Red lines show the mean exponentiated linear combination of periodic or trend or both features through the respective parameters. Dashed lines show random samples thereof; the shaded region marks the 25%-75% quantile.

For campylobacteriosis and, to a lesser extent, rotavirus reported incidence rates are higher in regions formerly belonging to eastern Germany (see Fig. 6.5). The parameters inferred for demographic components (see Fig. 6.5) show the role that age stratification plays for susceptibility. For all three diseases, a larger share of children and adolescents (ages 5-20 years) in the general population is indicative of increased incidence rates. Additionally, working-age adults (ages 20-65 years) appear to increase the incidence rate of borreliosis. It should be noted that this does not necessarily imply an increased susceptibility of the respective groups themselves, but could instead be due to latent variables correlated with age stratification, such as economic or cultural differences. The pairwise joint distributions reveal strong (anti-)correlations of the coefficients associated with the demographic and political components. E.g. the coefficient associated with age group [20-65) is strongly correlated with the coefficient associated with the east/west component, which implies ambiguity in the optimal choice of parameters.

The posterior probability over the dispersion parameter α (see Fig.6.5) is tightly distributed around the respective disease specific means. With small values of α, the distribution of case counts for campylobacteriosis approaches a Poisson

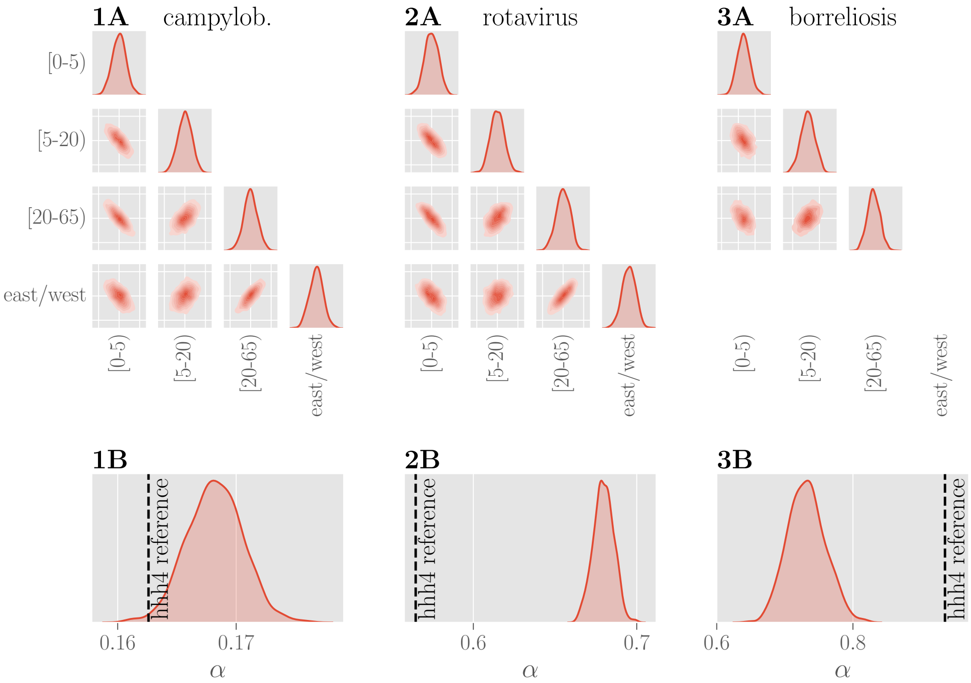


Figure 6.5. Learned weights for political and demographic components. Plots of the pairwise marginal distributions between inferred coefficients for three age groups and the east/west component for all three diseases are shown in **row A**. The marginal distribution of each coefficient shows a narrow unimodal peak, yet the pairwise distributions show that the individual features are clearly not independent. **Row B** shows the inferred posterior distributions of the overdispersion parameter α for three diseases. Values of α obtained using the *hhh4* reference model are indicated with a dashed black line. The inferred values for the dispersion parameter α are different, yet of similar magnitude, between the two models.

distribution, whereas the corresponding distributions for rotavirus and borreliosis are over-dispersed and deviate more from Poisson distributions.

6.4.2 Predictive performance

The one-week-ahead predictions are shown in Fig.6.6, for two selected cities (Dortmund and Leipzig for campylobacteriosis and rotavirus, Nürnberg (Nuremberg) and München (Munich) for borreliosis), together with the corresponding prediction from the reference *hhh4* model [217] fitted to the same data. A choropleth map of Germany (or the federal state of Bavaria in the case of borreliosis) shows the individual predictions for each county in one calendar week as an example. See also supplementary figures S8C Fig, S9C Fig and S10C Fig for predictions for 25 additional counties.

The BSTIM fits the mean of the underlying distribution of the data well. For rotavirus and borreliosis, it appears to overestimate the dispersion for the cities shown in Fig.6.6 as indicated by most data points falling within the inner

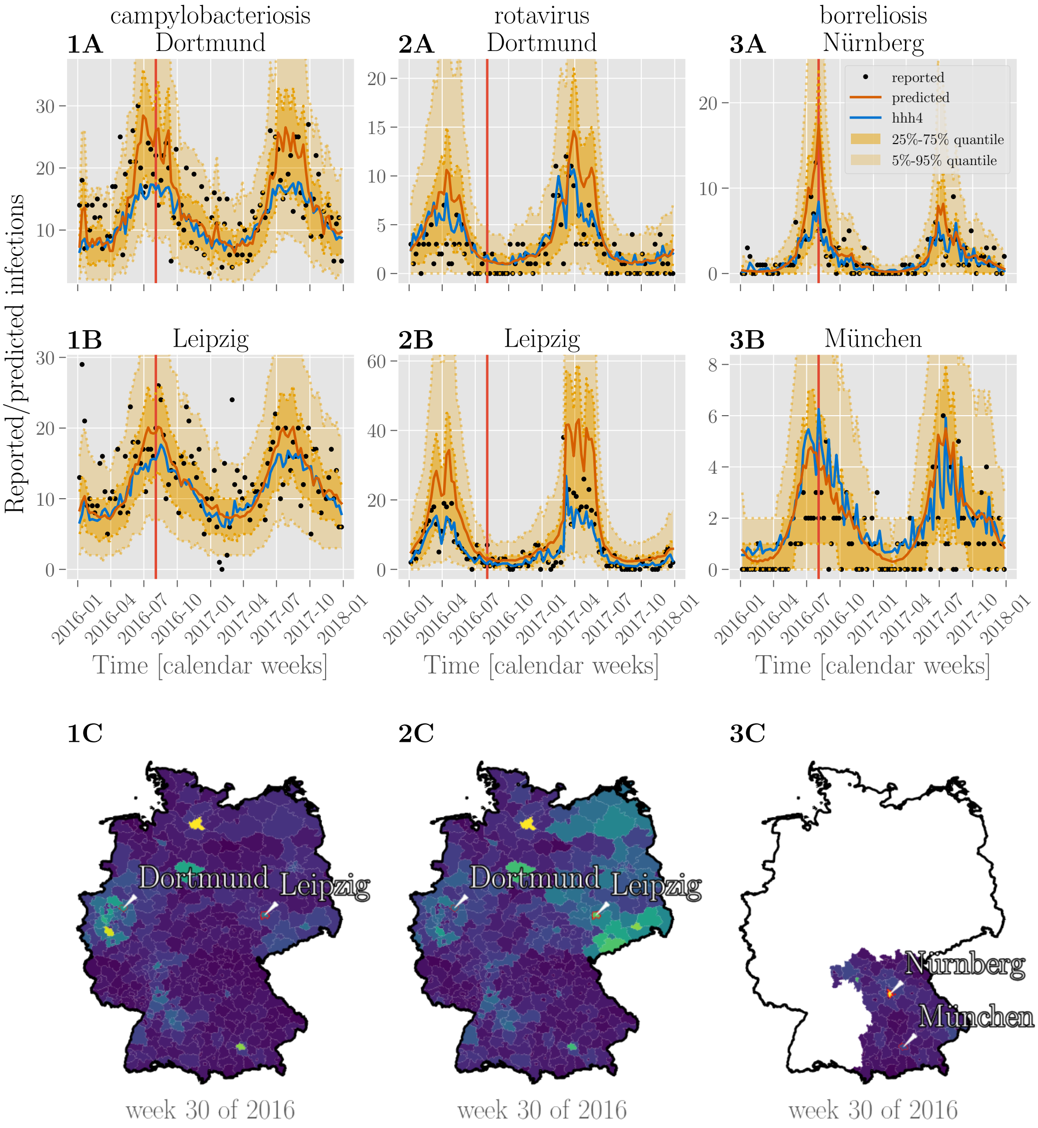


Figure 6.6. Predictions of case counts for various diseases by county. Reported infections (black dots), predictions of case counts by BSTIM (orange line) and the *hhh4* reference model (blue line) for campylobacteriosis **(column 1)**, rotavirus **(column 2)** and borreliosis **(column 3)** for two counties in Germany (for campylobacteriosis and rotavirus) or Bavaria (borreliosis), are shown in **rows A and B**. The shaded areas show the inner 25%-75% and 5%-95% percentile. **Row C** shows predictions of the respective disease for each county in Germany or the federal state of Bavaria in week 30 of 2016 (indicated by a vertical red line in rows A and B). Information about the shape of counties within Germany is publicly provided by the German federal agency for cartography and geodesy (Bundesamt für Kartographie und Geodäsie) (GeoBasis-DE / BKG 2018) under the dl-de/by-2-0 license.

25%-75% quantile. This may be due to a too high dispersion parameter α (cf. Fig.6.5) or uncertainty about model parameters. It should be noted, however, that the optimal dispersion parameter itself may vary from county to county, whereas our model infers only one single value for all counties together. The resulting predictions for all three diseases are smoother in time and space (cf. the chloropleth maps in Fig.6.6) than the predictions by the reference *hhh4* model. We attribute this to the smooth temporal basis functions and spatio-temporal interaction kernel of our model.

To quantitatively compare the performance of both models, we calculate the distributions of deviance and Dawid-Sebastiani score over all counties for BSTIM and the *hhh4* reference model as shown in Fig.6.7. Both measures show a very similar distribution of errors between both models for all three diseases, as it can be seen in table 6.2. Only for borreliosis, the *hhh4* model appears to be more sensitive to outliers.

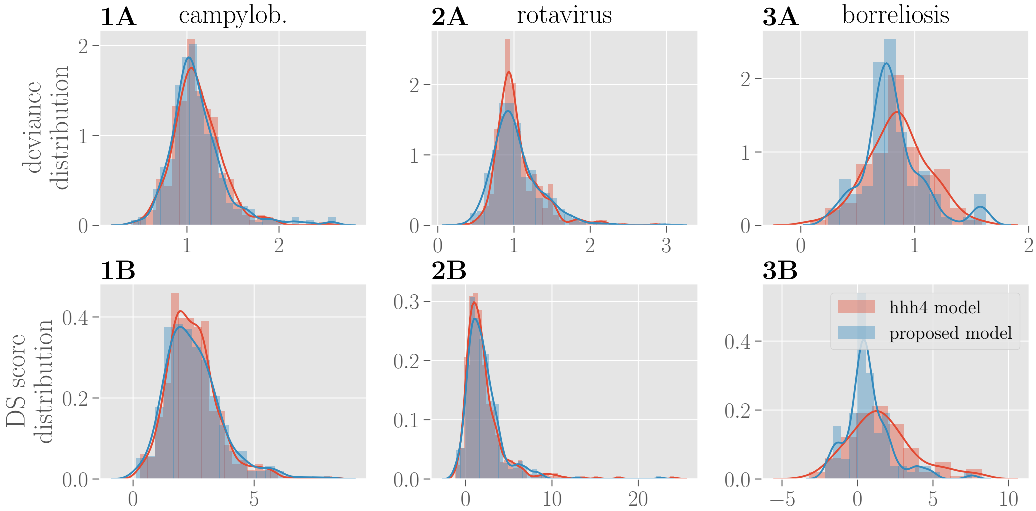


Figure 6.7. Evaluation of prediction performance. The distribution of deviance over counties is shown in **row A** for BSTIM (blue) and the reference *hhh4* model (red) for all three diseases. The corresponding distribution of Dawid-Sebastiani scores is shown in **row B**.

disease	score	BSTIM	*hhh4*
campylob.	deviance	1.11 ± 0.3	1.11 ± 0.26
	DS score	2.49 ± 1.17	2.47 ± 1.06
rotavirus	deviance	1.03 ± 0.32	1.04 ± 0.3
	DS score	2.08 ± 2.17	2.1 ± 2.54
borreliosis	deviance	0.81 ± 0.27	0.85 ± 0.27
	DS score	0.74 ± 1.54	1.63 ± 2.24

Table 6.2. Deviance and Dawid-Sebastiani score (mean ± standard deviation) for all three diseases and both BSTIM and the *hhh4* model.

6.4.3 Benefits of probabilistic modeling for epidemiology

Probabilistic modeling relies on the specification of prior probability distributions over parameters [190]. In the context of epidemiology, this makes it possible to incorporate domain knowledge (e.g. we know that case counts tend to be overdispersed relative to Poisson distributions, but not to which degree for a specific disease) as well as modeling assumptions. This is particularly relevant for diseases with limited available data (e.g. those not routinely recorded through surveillance), where appropriately chosen priors are required to prevent overfitting. The framework can easily be extended to include additional features or latent variables. For example, we introduce precise locations and times of individual cases as latent variables, given only the counties and calendar weeks in which they occurred.

Probabilistic models as discussed here provide samples of the posterior distribution of parameters as well as model predictions. This allows for analysis that is not possible with point estimation techniques such as maximum likelihood estimation. In epidemiology, datasets can be small, noisy or collected with low spatial or temporal resolution. This can lead to ambiguity, where the observations could be equally well attributed to different features and thus different model parameterizations are plausible. While maximum likelihood estimation in such a situation selects only the single most likely model, Bayesian modeling captures the full distribution over possible parameters and predictions, and thus preserves information about the uncertainty associated with the parameters of the model itself. Analyzing the parameter distribution can thus help identify redundant or uninformative features. For example, an inspection of the posterior marginal distributions of the model parameters in the supplementary figure S1C Fig shows, that e.g. the first parameter associated with the trend component, that constitutes an additive “bias” term, is subject to larger variance, which could indicate, that this coefficient is redundant given the other features and might inform further investigation.

Samples from the inferred parameter distributions are afterwards used to derive samples of predicted future cases. The resulting predictions thus incorporate both noise assumptions about the data as well as model uncertainty. This can be relevant for determining confidence intervals, in particular in situations where model uncertainty is large. The samples of the predictive distribution can in turn be used for additional processing, or if predictions in the form of point estimates are desired, they can be summarized by the posterior mean.

6.4.4 Possible extensions

To account for overdispersion in the data, we use a Negative Binomial distribution in this study. Other choices are possible, e.g. zero-inflated distributions [197, 201] or quasi-Poisson distributions [218], each of which has a different implication for the resulting model. Since the Negative Binomial distribution assigns more weight to smaller counts relative to quasi-Poisson [218], the latter may be a more adequate choice when accurately predicting higher counts, e.g. during outbreaks, is critical. If there are differences between individual counties, that are suspected to lead to varying degrees of overdispersion, the overdispersion parameter α of the Negative Binomial distribution could also be chosen to vary in time and space like the corresponding mean μ [219, 220].

Whereas spatio-temporal interaction effects are here modeled as a function of geographical proximity, the kernel's composite basis functions make it possible to use alternative spatial distance measures, e.g. derived from transportation networks for people or goods [221]. For diseases where the kernel clearly factorizes into a single temporal and spatial component, a simpler spatial kernel function with a parameter for the bandwidth could be chosen. This allows including further prior assumptions or constraints, e.g. strict non-negativity or power law characteristics of interactions [222].

Due to the flexibility of the probabilistic modeling and sampling approach, additional variables can be easily included and their influence analyzed (e.g. weather data, geographical features like forests, mountains and water bodies, the location and size of hospitals, vaccination rates, migration statistics, socioeconomic features, population densities, self-reported infections on social media [172], work, school and national holidays, weekends and large public events). For features where precise values are not known, probability distributions could be specified and included in the probabilistic model, which could improve the model's estimate of uncertainty. For example, since the precise locations and times of individual infections are not publicly known, we simply assume a geographically and temporally uniform distribution of cases within the given county and calendar week. The conditional probability distributions could be refined by incorporating additional information (e.g. weekends and population density maps). However, precise information about place and time of infections are available to local health agencies. The model presented here could readily be implemented there to use this more accurate data.

In this study, we assume that the presented model, due to time-varying features as well as interaction effects is flexible enough to model the dynamics of the diseases in question throughout the year. There may, however, be influential latent variables that cannot be explicitly included as exogenous variables, in particular for diseases with very pronounced epidemic outbreaks. In such cases, the 'outbreak' stage of the disease could be modeled separately from the baseline stage, thereby increasing the degrees of freedom in the model. This has been demonstrated for dengue fever [197], where Markov switching is employed to detect sudden changes in the expected number of cases and provide early warnings when such a state transition occurs.

6.5 Conclusion

In this paper, a probabilistic model is proposed for predicting case counts of epidemic diseases. It takes into account a history of reported cases in a spatially extended region and employs MCMC sampling techniques to derive posterior parameter distributions, which in turn are incorporated in predicted probability distributions of future infection counts across time and space.

For all three tested diseases (campylobacteriosis, rotavirus and borreliosis) the same model, using interaction effects, temporal, political and demographic information, performs well and produces smooth predictions in time and space. For each disease, the inferred spatio-temporal kernels capture the specific interaction effects in a single function, that can be visualized and interpreted, and can be applied regardless of the topology of counties or their neighborhood relationships. A comparison with the standard *hhh4* model, which uses maximum likelihood estimation instead of Bayesian inference, shows comparable performance. At the expense of higher computational costs than the point estimate used in *hhh4*, the sampling approach employed here provides information about the full posterior distribution of parameters and predictions. The posterior parameter distribution provides information about the relevance of the corresponding features for the inferred model, and helps in identifying redundant features or violated model assumptions. The inferred features of our model are interpretable and their individual contribution to the model prediction can be analyzed: spatio-temporal interactions reveal information about the dynamic spread of the disease, temporal features capture seasonal fluctuations and long-term trends, and the assigned weights indicate relevance of additional features. The posterior predictive distribution also accounts for the uncertainty about parameters, e.g. due to simplifying model assumptions or a lack of data, rather than just the variability inherent in the data itself. This additional information is valuable for public-health policy-making, where accurate quantification of uncertainty is critical.

Acknowledgments

We would like to thank our editor and anonymous reviewers for their comments which helped us to improve the paper.

Chapter 7

Example 3: predicting environmental variables

Environmental scientists often have to reason about complex phenomena influenced by many factors. One of the typical applications of remote sensing is predicting the leaf area index (LAI). LAI is a unitless measure ($\mathrm{m}^2\,\mathrm{m}^{-2}$) of a plant's one-sided leaf surface area relative to the soil surface area. [223] Since it characterizes the photosynthetic performance of plants, the size and density of the crop's canopy, it is a good indicator of a plant's growth, health, and stand productivity. [224, 225, 223] It is, therefore, often used in modeling in agriculture and environmental sciences. [225, 226, 227, 228] LAI is influenced by many factors, like the type of soil or crop, weather conditions (e.g., the amount of precipitation, sunshine duration), etc. [229, 230] These factors vary throughout the crop's life cycle, meaning that environmental scientists have to measure other variables simultaneously, over a long period of time. The measurements are often carried out in different locations or climate zones. This results in heterogeneous data, and there might be few measurements for a particular location or year. Therefore, an algorithm for the estimation of environmental variables has to be able to deal with heterogeneous data.

We developed an interpretable model for predicting the leaf area index of white winter wheat for which we used simultaneous measurements of LAI and reflectance spectra of white winter wheat in two locations over four years. Here, the biggest challenge was working with data with such systematic differences while preserving the important information from measurements to make an accurate predictive model. Besides accounting for systematic differences in data, we also wanted to include domain knowledge in the model to understand how measured variables (in our case, different wavelengths of the recorded spectra) affect predictions of LAI. Given those reasons, we chose a similar Bayesian approach as in chapter *Example 2: predicting infectious diseases.*

However, different from the prediction of infectious diseases, we chose a hierarchical model to predict LAI. Bayesian hierarchical models are a class of models that allow parameters to vary on multiple levels of abstraction. In our case, it means the model parameters for different fields and years can vary around a common value, which preserves similarities among the groups of measurements, while allowing for specific differences between the datasets. This property makes Bayesian hierarchical models particularly suitable for environmental sciences [231], where they have been used to assess many phenomena such as the effects of climate change through land surface phenology [232, 233, 234, 235] or for estimating indicators such as Normalized Difference Vegetation Index [236].

We first modeled reflectance spectra using spline basis functions with adaptive knot placement. [237] This captures the association between LAI and the spectral reflectance at various wavelengths. We then developed three different Bayesian hierarchical generalized linear models of varying complexity. The first model is a baseline model where all data is pooled together, and there are no distinctions between fields and years of measurement. The second model adds a bias parameter for each dataset, which accounts for the variation in scale between the datasets. The third model further adds a bias parameter for each year. The model infers a full posterior distribution over model parameters for each model to provide a full posterior predictive distribution over LAI values. We again use MCMC sampling implemented in the probabilistic programming software package PyMC3 [190] to infer posterior distributions of all model parameters and predictions. We further compute feature importance (see chapter *What is interpretable machine learning*) as a causal method to break dependence between correlated features and calculate the relative change in the model's error before and after breaking that dependence. In this way, we can evaluate the contribution that our model assigns to each feature of modeled reflectance spectra.

In general, each model has a relatively good performance of LAI predictions. In addition to predicting LAI, all three models learn an interpretable kernel-like function of reflectance spectra. The main difference between the models is their complexity and how well the inferred kernel function describes the model. With feature importance we can see how parts of reflectance spectra (which correspond to physical phenomena) contribute to predictions, which makes the model globally interpretable on a modular level. Our analysis showed that adding a bias parameter for each dataset greatly improves the model, while adding a bias parameter for each year increases the complexity of the model with little additional gain.

Chapter 8

Bayesian hierarchical models can infer interpretable predictions of leaf area index from heterogeneous datasets

Contributions

This chapter was first published as: *Stojanović O, Siegmann B, Jarmer T, Pipa G, Leugering J (2022) Bayesian Hierarchical Models can Infer Interpretable Predictions of Leaf Area Index From Heterogeneous Datasets. Frontiers in Environmental Science, 717. https://doi.org/10.3389/fenvs.2021.780814*

Olivera Stojanović
Roles: Conceptualization, Methodology, Project administration, Software, Visualization, Writing — original draft
Affiliation: Neuroinformatics Lab, Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany

Bastian Siegmann
Roles: Data curation, Conceptualization, Validation, Writing — review and editing
Affiliation: Institute of Bio- and Geosciences, Plant Sciences (IBG-2), Jülich Research Centre, Jülich, Germany

Thomas Jarmer
Roles: Data curation, Conceptualization, Validation, Writing — review and editing
Affiliation: Working Group Remote Sensing and Digital Image Analysis, Institute of Computer Science, University of Osnabrück, Osnabrück, Germany

Gordon Pipa
Roles: Supervision, Resources, Conceptualization, Validation, Writing — review and editing

Affiliation: Neuroinformatics Lab, Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany

Johannes Leugering
Roles: Conceptualization, Methodology, Project administration, Software, Visualization, Writing — original draft
Affiliation: Neuroinformatics Lab, Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany

Data Availability Statement

The datasets generated for this study and the corresponding source code can be found online at: `https://github.com/ostojanovic/bayesian_lai`.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8.1 Abstract

Environmental scientists often face the challenge of predicting a complex phenomenon from a heterogeneous collection of datasets that exhibit systematic differences. Accounting for these differences usually requires including additional parameters in the predictive models, which increases the probability of overfitting, particularly on small datasets. We investigate how Bayesian hierarchical models can help mitigate this problem by allowing the practitioner to incorporate information about the structure of the dataset explicitly. To this end, we look at a typical application in remote sensing: the estimation of leaf area index of white winter wheat, an important indicator for agronomical modeling, using measurements of reflectance spectra collected at different locations and growth stages. Since the insights gained from such a model could be used to inform policy or business decisions, the interpretability of the model is a primary concern. We, therefore, focus on models that capture the association between leaf area index and the spectral reflectance at various wavelengths by spline-based kernel functions, which can be visually inspected and analyzed. We compare models with three different levels of hierarchy: a non-hierarchical baseline model, a model with hierarchical bias parameter, and a model in which bias and kernel parameters are hierarchically structured. We analyze them using Markov Chain Monte Carlo sampling diagnostics and an intervention-based measure of feature importance. The improved robustness and interpretability of this approach show that Bayesian hierarchical models are a versatile tool for the prediction of leaf area index, particularly in scenarios where the available data sources are heterogeneous.

8.2 Introduction

The leaf area index (LAI) is a unitless measure ($\mathrm{m}^2\,\mathrm{m}^{-2}$) of the one-sided leaf surface area of a plant relative to the soil surface area. [223] It characterizes, among other variables, the photosynthetic performance of plants [224, 225, 223], the size and density of the crop's canopy and thus serves as an important indicator for the plant's growth stage and stand productivity [238, 226, 227, 228, 239]. It plays a major role in meteorological, ecological, and agronomical modeling [240, 241, 242, 243, 244, 245], as well as for studying the influence of climate change on crop growth [246, 247].

Various non-destructive methods exist to measure or estimate LAI directly [248], but they typically require taking a large number of manual measurements in the field. Since this is a laborious process and it can be difficult to control for confounding variables such as weather, alternative faster approaches to infer LAI from indirect measures, e.g., spectroscopy and (hyper-)spectral imaging, have been investigated [226, 249, 250, 230]. A common approach makes use of vegetation indices (VIs), which can be computed from distinct wavebands of spectral measurements, to estimate LAI [251]. These measures, while well understood and easy to calculate, have several limitations. For example, most of them are sensitive to more than one plant parameter (e.g., LAI and chlorophyll content) [252, 253], and especially for wheat crops, the non-linear relationship between numerous VIs and the LAI can lead to saturation for moderate to high LAI values (LAI > 3) [254, 255]. We instead use a Bayesian, spline-based regression method that utilizes the entire hyperspectral reflectance measurement to predict LAI and provides uncertainty estimates over all model parameters.

However, the relationship between LAI and spectral reflectance is also affected by other factors, such as the crop type, phenology, sun illumination, local micro-climate, the type of soil, or the amount of precipitation [229, 230], and it may vary throughout the life cycle of the crop. These effects can be included in spatio-temporal models of LAI [256, 257, 234, 235, 258], which can be applied to data from aerial or satellite surveillance. This has the potential to greatly simplify monitoring crop growth across large or remote areas [259, 260, 261].

But the annotated training data required for such spatio-temporal models, i.e., matched measurements of reflectance spectra, ground-truth LAI, and potential confounding variables, is not available for every location and crop. For example, the data used in this study consists of distinct datasets of corresponding LAI measurements and reflectance spectra, each set acquired on a single field over a period of one or two days. In this situation, the amount of labeled training data that can be acquired is too limited to fit either a full spatio-temporal model or a separate model for each field and/or growth stage. It is possible, in principle, to train a single model that generalizes across these conditions by simply pooling multiple datasets that were acquired under different conditions (see [262], for example). However, since the association between each data point and the specific dataset it belongs to (and thus its location and time) is lost in the pooling process, such a model is likely to perform worse than a spatio-temporal model or a field- and growth-stage-specific model, given sufficient training data. Ideally, we would like to find a compromise between these two extremes, i.e., between a single model trained on the pooled data on the one hand and an independent model for each dataset, on the other hand, that allows us to generalize over all the available datasets yet makes specifically adjusted predictions for each dataset. To this end, we propose a hierarchical, parameter-efficient Bayesian model, which implicitly accounts for the influence of location and time by allowing the model parameters to vary across different datasets.

Bayesian hierarchical models similar to the one suggested here are especially appealing for environmental sciences [231], where they have seen increasingly widespread use. For example, several recent studies applied Bayesian hierarchical models to time series of multispectral satellite images in order to assess the effects of climate change through land surface phenology [232, 233, 234, 235] or other indicators such as Normalized Difference Vegetation Index [236]. A similar remote sensing approach has been used to predict LAI and its spatio-temporal evolution for bamboo [263] and other forests [264, 265]. For agronomical models of the LAI of food crops such as rice [266], Brazilian Cowpea [267] or white winter wheat [268], local multispectral measurements are often used instead of – or in addition to – satellite images. In most of these studies, Bayesian hierarchical models are used to impose prior domain knowledge, combine multimodal data sources, and integrate data collected at multiple resolutions of space and/or time, all of which ultimately improve prediction performance. By contrast, our primary goal is to show how Bayesian hierarchical models and associated tools can be used to construct and diagnose simple and interpretable models for heterogeneous datasets, which commonly occur in environmental sciences.

Based on these considerations, we develop a Bayesian hierarchical model according to the following steps: (1) we filter the spectral measurements to remove noise, (2) apply basis splines with adaptively placed knots to extract features from the spectra, (3) train a Bayesian hierarchical model to predict LAI from these features on labeled data, (4) select and validate the best performing model, (5) and estimate the importance of the individual features for prediction. Our model learns an easily interpretable general relationship between reflectance spectra and LAI, as well as the dataset-specific deviations from that baseline. By using a variant of Markov Chain Monte Carlo (MCMC) sampling, we can incorporate domain knowledge or regularization through prior distributions of the parameters and provide a full posterior probability distribution over these parameters, which allows the quantification of uncertainty. We compare two variants of this hierarchical approach with a non-hierarchical alternative and find that it indeed offers a favorable trade-off between prediction accuracy and model complexity.

8.3 Methods

8.3.1 The dataset

We evaluate our proposed model on a combination of four datasets, totaling 191 pairs of measured reflectance spectra (see also Supplementary Material, Figure S1D for examples) and corresponding measurements of the LAI on fields covered by white winter wheat (lat. *Triticum aestivum*).

Each pair of measurements was taken on a different square plot of size 50 cm × 50 cm. The LAI values of each plot were measured multiple times in a non-destructive way and averaged to a single value per plot. Five reflectance spectra were acquired and averaged for each plot using a spectroradiometer from a height of 1.4 m above ground with a nadir view and converted to absolute reflectance values using a reflectance standard of known reflectivity (Spectralon, Labsphere Inc., USA). The data were collected on four different fields in different years, corresponding to different stages in the plants' growth cycle, and there are minor differences in the data collection procedure.

The first two sampling areas, which we call Field A and Field B in the following, are located near Köthen, Germany, which is a part of one of the most important agricultural regions in Germany. The region is distinctly dry, with 430 mm mean annual precipitation due to its location in the Harz mountains. The mean annual temperature varies between 8 °C to 9 °C. The study area has an altitude of 70 m above sea level and is characterized by a Loess layer up to 1.2 m deep that covers a slightly undulated tertiary plain. The predominant soil types of the region are Chernozems, in conjunction with Cambisols and Luvisols. At two locations in this region, 57 spectral measurements were recorded on 7th to 8th May 2011 using two ASD FieldSpec III spectroradiometers (ASD Inc., USA) with 25° field of view optics. Another 74 measurements were taken on 24th to 25th May 2012, using one ASD FieldSpec III (ASD Inc., USA) and one SVC HR-1024 spectroradiometer (Spectra Vista Corporation, USA) with 14° field of view. For each location, the corresponding LAI was measured non-destructively with a SunScan device (Delta-T Devices Ltd., USA) in 2011 and an LAI 2000 (LI-COR Inc., USA) in 2012, and, respectively, five and four LAI measurements were averaged per plot. Data from this study area was also used and described in more detail in [262].

The other two sampling areas, called Field C and Field D in the following, are located near Demmin, Germany. The region has a mean annual precipitation of 550 mm, and a mean annual temperature of 8 °C. Albeluvisols interspersed by Haplic Luvisols dominate the sand-rich area. The observed area is south of the river Tollense, where the ground elevation drops from 70 m to 7 m due to glacial moraines causing high variability in soil conditions. At two locations in this region, 26 spectral measurements were recorded on 5th June 2015, and another 34 on 10th May 2016 using SVC HR-1024i spectroradiometer (Spectra Vista Corporation, USA) in nadir view 1.4 m above the ground using 14° field of view optics. In this case, six measurements of LAI (taken with LAI 2000 (LI-COR Inc., USA)) were averaged for each plot.

The recorded spectra cover wavelengths in the range from 350 nm to 2500 nm, of which we use the range from 400 nm to 1350 nm (for details see table 8.1). We preprocess these spectra by smoothing with a first-order Gaussian filter with width $\sigma = 10\,\mathrm{nm}$.

For a summary of these parameters, see table 8.1.

In the following, we reference specific subsets of this data. We introduce the following notation: all spectra-LAI-pairs for field $j \in \{1, \ldots, 4\}$ are numerically indexed by the set J^j, where J^1, J^2, J^3, J^4 correspond to the measurements from

	Field A	Field B	Field C	Field D
collection date	7th to 8th May 2011	24th to 25th May 2012	5th June 2015	10th May 2016
measurements	57	74	26	34
location	Köthen, Germany		Demmin, Germany	
LAI device	SS1 SunScan	LAI-2000	LAI-2000	
spectral device	ASD FieldSpec III	SVC HR-1024	SVC HR-1024i	
field of view	25°	14°	14°	
measurement height	1.4 m above ground			
reflectance standard	Spectralon, Labsphere Inc., USA			
spectral range measured	350 nm to 2500 nm			
spectral range used	400 nm to 1350 nm at a resolution of 1 nm, smoothing with $\sigma = 10\,\mathrm{nm}$			
LAI range	0.5 to 3.32	1.14 to 6.16	1.72 to 7.46	0.48 to 5.25

Table 8.1. Parameters of the dataset and the collection procedures.

Field A, Field B, Field C, and Field D, respectively. We denote the $i^{\text{th}} \in J^j$ reflectance spectrum from dataset j by the function $R_i^j(\lambda)$ of the wavelength $\lambda \in [400\,\mathrm{nm}, 1350\,\mathrm{nm}]$, and the corresponding measured LAI value by Y_i^j.

8.3.2 Feature extraction from reflectance spectra with a spline basis

The data collection and preprocessing steps outlined above result in reflectance spectra of wavelengths 400 nm to 1350 nm at a resolution of 1 nm. Since this representation is much too high dimensional for direct use, we extract the most important information into a low dimensional representation by computing the inner product between the preprocessed reflectance spectra and a set of eleven cubic basis splines (B-splines) with adaptively placed knots (see figure 8.1).

The positions $\kappa_i, i \in \{1, \dots, 11\}$ of the inner knots, which determine the shape of the individual basis splines, are chosen such that the cumulative absolute curvature $Q(\kappa_{i+1}) - Q(\kappa_i)$ of the average reflectance spectrum is equal between any two successive knots i and $i+1$. We compute the absolute curvature $q(\lambda)$ by convolving the average reflectance spectrum $\bar{R}$ with the second derivative of the Gaussian function g, and then compute the absolute value thereof[1]. Formally, we can express this as follows:

[1] This is equivalent to computing the absolute curvature of the smoothed average reflectance spectrum $g * \bar{R}$.

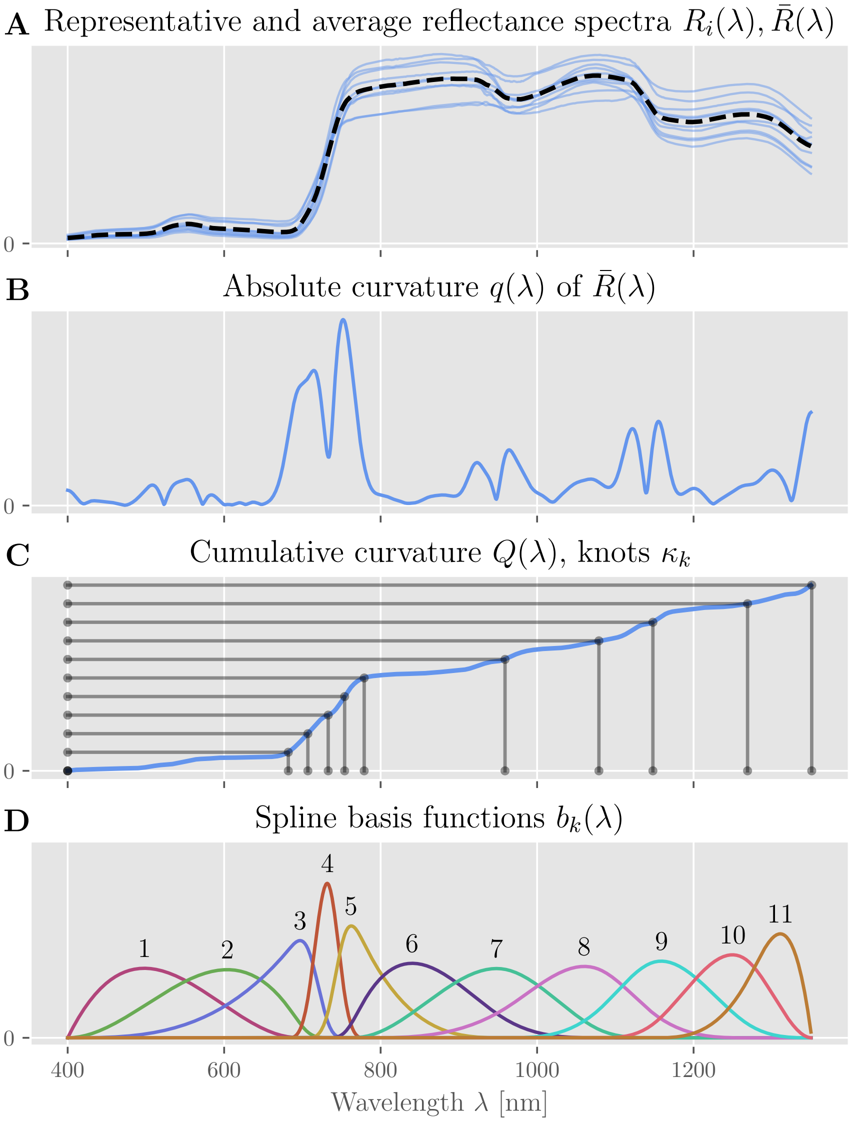


Figure 8.1. Adaptive knot-placement for B-Splines. (**A**) For the measured reflectance spectra $R_i(\lambda)$, (**B**) we calculate the mean absolute curvature $q(\lambda)$. (**C**) We then find knot positions such that the integral $Q(\lambda)$ of this measure between any two successive knots κ_i, κ_{i+1} is identical. (**D**) The result are 11 cubic spline basis functions $b_k(\lambda)$ with non-uniformly spaced knots.

$$
\begin{aligned}
g(\lambda) &= \frac{1\,\mathrm{nm}}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\lambda^2}{\sigma^2}\right), && \sigma = 10\,\mathrm{nm}\\
\bar{R}(\lambda) &= \frac{1}{\sum_{j=1}^{4}|J^j|}\sum_{j=1}^{4}\sum_{i\in J^j} R_i^j(\lambda)\\
q(\lambda) &= \left|(\bar{R} * g'')(\lambda)\right|\\
Q(x) &= \int_{400\,\mathrm{nm}}^{x} q(\lambda)\mathrm{d}\lambda && \forall x \in [400\,\mathrm{nm}, 1350\,\mathrm{nm}]\\
Q_{\max} &= Q(1350\,\mathrm{nm})\\
\kappa_i &= Q^{-1}\left(\frac{i-1}{10}Q_{\max}\right) && \forall i \in \{1, \ldots, 11\}
\end{aligned}
$$

The eleven basis functions $b_k, k \in \{1, \ldots, 11\}$ are generated from this knot vector κ using the standard Cox-DeBoor algorithm [269], where the multiplicity of the first and last knot is three, i.e., all basis functions go to zero at their respective start and end knots. The last basis function b_{12}, which originates at knot κ_{10}, is not used. This heuristic algorithm results in a proportionally larger number of knots, and thus higher spatial resolution, where the reflectance spectra have the largest absolute curvature and hence "have most structure"; see also [237, 270, 271]. For each of the data-subsets $j \in \{1, \ldots, 4\}$, we can then compute our model's *feature* or *design matrix* X^j using these basis functions $b_k(\lambda)$:

$$
(X^j)_{i,k} = \langle R_i^j, b_k\rangle = \int_{400\,\mathrm{nm}}^{1350\,\mathrm{nm}} R_i^j(\lambda)b_k(\lambda)\mathrm{d}\lambda, \quad \forall i \in J^j, k \in \{1, \ldots, 11\} \tag{8.1}
$$

8.3.3 Bayesian Markov Chain Monte Carlo regression for predicting LAI

Our primary objective is to construct a simple, interpretable model that can reliably predict the LAI value of a wheat plot directly from a corresponding reflectance spectrum. We are additionally interested in analyzing the model's confidence, how well it generalizes, and which features it relies on most to make a prediction. Since the total available data is limited and stems from four heterogeneous datasets, prior constraints are required to prevent overfitting.

In order to meet all of these requirements, we design three different (non-)hierarchical Bayesian generalized linear models (GLM) [272, 273] of different complexity. For each of these, we infer a full posterior distribution over model parameters from training data and use this to provide a full posterior predictive distribution over LAI scores on testing data. To generate representative samples from these probability distributions, we use a specific type of Hamiltonian Monte Carlo sampling, namely No-U-Turn-Sampling [205] (NUTS), as implemented by the probabilistic programming package *pyMC3* [190].

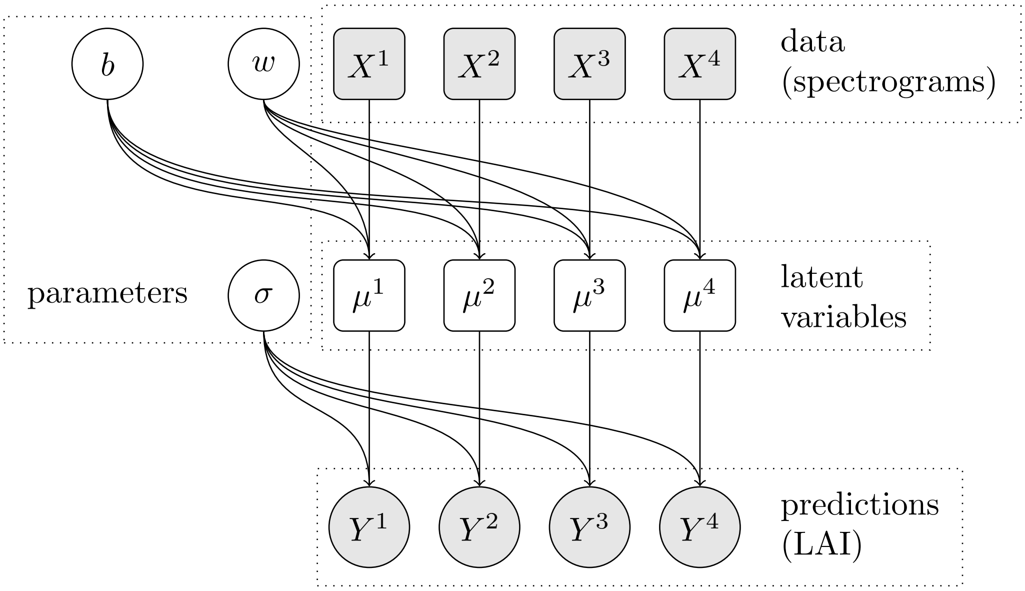


Figure 8.2. Dependency graph of the baseline model. For each dataset j (encoded in the feature matrix X^j and corresponding labels Y^j), the prediction depends on the same three shared parameters b, w and σ. Circles represent random variables, rectangles represent deterministic variables, filled shapes represent observed variables.

Model 1: A baseline model with pooled data

As a baseline (see figure 8.2), we construct a simple generalized linear model, which we apply to all of the datasets $j \in \{1, \dots 4\}$ together. This model merely pools all available data but does not account for any systematic differences that might exist between the individual datasets. We assume the logarithm of the observed LAI scores to be normally distributed around an affine linear predictor μ^j with deviation σ, which is a model parameter with log-normal prior. The predictor μ^j is computed by the matrix-vector product between the dataset's feature matrix X^j and the model's weight vector $w = (w_1, \dots, w_{11})$, plus an additional bias parameter b. Including the unknown deviation parameter σ, the model thus has a total of 13 free parameters to be inferred from data. The individual parameters w_k and b have normal priors with standard deviation 1 and 11, respectively, to allow the individual bias term to counteract the effect of all 11 weights, if necessary. The baseline model is described by equation 8.2.

$$\begin{aligned}
\log(\sigma) &\sim \text{Normal}(0,1) & \\
b &\sim \text{Normal}(0,11) & \\
w_k &\sim \text{Normal}(0,1) & \forall k \in \{1,\dots,11\} \\
\mu^j &= X^j w + b, & \forall j \in \{1,\dots,4\} \\
\log(Y_i^j) &\sim \text{Normal}(\mu_i^j, \sigma), & \forall j \in \{1,\dots,4\}, i \in I^j
\end{aligned} \tag{8.2}$$

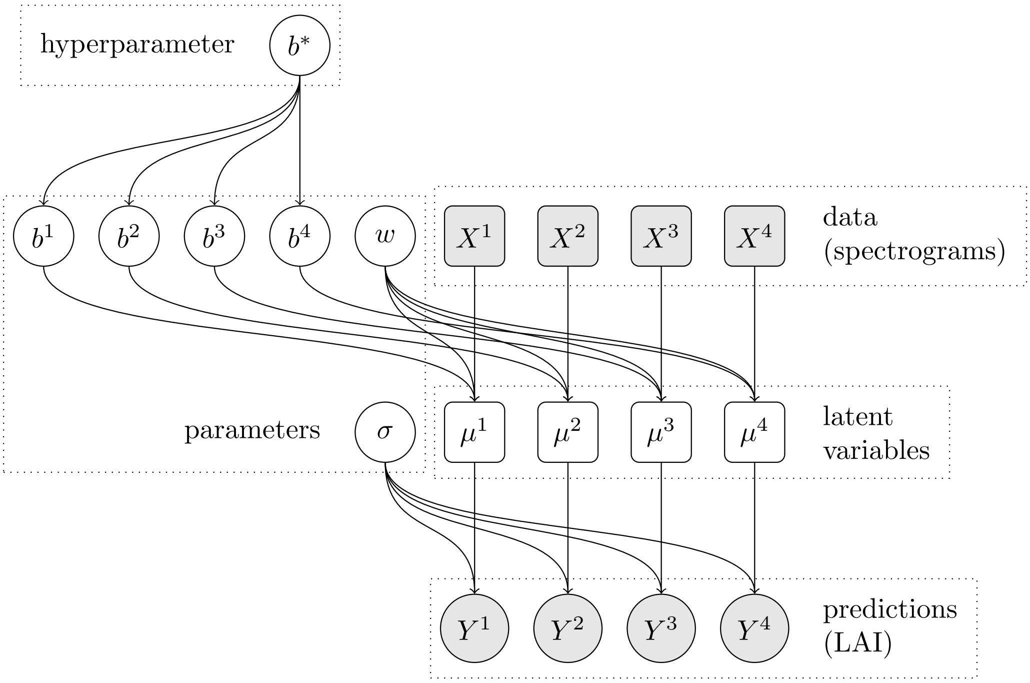


Figure 8.3. Dependency graph of the hierarchical model with dataset-specific bias terms. The predictions for each dataset j depend on an individual bias parameter b^j, which in turn depends on the shared mean bias parameter b^*.

Model 2: A model with hierarchical bias

Our second model (see figure 8.3) extends the baseline model by an additional bias parameter b^j for each dataset and thus has a total of 17 free parameters. Due to the logarithmic link function, this additional parameter per dataset allows accounting for the overall variation in scale between the four different datasets. But rather than setting each parameter b^j independently (and thus adding three full degrees of freedom), we constrain them to be clustered around a common bias value b^*, which replaces the bias term b in the baseline model. Therefore, the prior for the new variables b^j is a Normal distribution centered at b^* with an order of magnitude smaller standard deviation of $11/10 = 1.1$. The affine linear predictor μ^j then depends on the dataset-specific bias term b^j. The hierarchical bias model is described by equation 8.3.

$$\begin{aligned}
\log(\sigma) &\sim \text{Normal}(0,1) \\
b^* &\sim \text{Normal}(0,11) \\
b^j &\sim \text{Normal}(b^*,1.1) && \forall j \in \{1,\dots,4\} \\
w_k &\sim \text{Normal}(0,1) && \forall k \in \{1,\dots,11\} \\
\mu^j &= X^j w + b^j, && \forall j \in \{1,\dots,4\} \\
\log(Y_i^j) &\sim \text{Normal}(\mu_i^j,\sigma), && \forall j \in \{1,\dots,4\}, i \in I^j
\end{aligned} \tag{8.3}$$

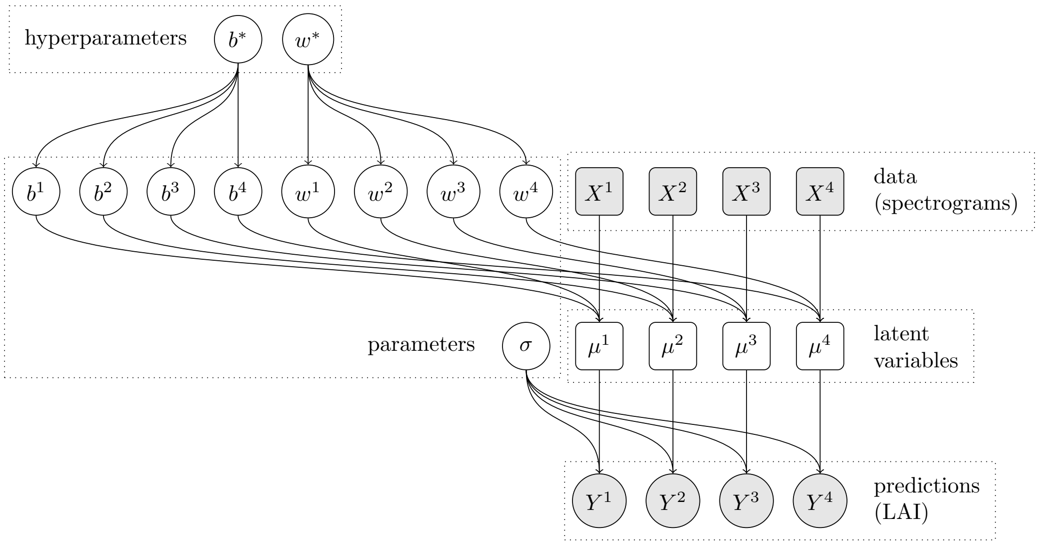


Figure 8.4. Dependency graph of the hierarchical model with dataset-specific bias and weight terms. The predictions for each dataset j now depend on an individual bias parameter b^j and weight vector w^j, which in turn depend on the shared bias parameter b^* and the shared weight vector w^*, respectively.

Model 3: Full hierarchical model

Our third model (see figure 8.4) extends the second model even further by also allowing the model weight vector w to vary for each dataset. Just like we did for the bias terms, we introduce the new parameter vectors w^j, and we constrain the individual parameters w_k^j to be clustered around the corresponding common values w_k^* with standard deviation 0.1. This increases the model's degrees of freedom by an additional 44 parameters (11 for each dataset), resulting in a total of 61 free parameters. The affine linear predictor μ^j then depends on a dataset-specific weight vector w^j and a dataset-specific bias term b^j. The full hierarchical model is described by equation 8.4.

$$\begin{aligned}
\log(\sigma) &\sim \text{Normal}(0,1) \\
b^* &\sim \text{Normal}(0,11) \\
b^j &\sim \text{Normal}(b^*,1.1) && \forall j \in \{1,\dots,4\} \\
w_k^* &\sim \text{Normal}(0,1) && \forall k \in \{1,\dots,11\} \\
w_k^j &\sim \text{Normal}(w_k^*,0.1) && \forall k \in \{1,\dots,11\}, j \in \{1,\dots,4\} \\
\mu^j &= X^j w^j + b^j, && \forall j \in \{1,\dots,4\} \\
\log(Y_i^j) &\sim \text{Normal}(\mu_i^j,\sigma), && \forall j \in \{1,\dots,4\}, i \in I^j
\end{aligned} \tag{8.4}$$

8.3.4 Model selection using Pareto-Smoothed Importance Sampling

To get an unbiased estimate of our model's generalization error from the very limited available data, we would like to perform leave-one-out cross-validation (LOO-CV) and compute the expected log posterior predictive density (ELPD) for new data. [274] Unfortunately, this is a prohibitively expensive computation when combined with MCMC sampling. However, the generated samples and their associated log-likelihood values contain sufficient information to estimate the LOO-CV ELPD by directly weighing the samples. This procedure is called Pareto-smoothed importance sampling (PSIS) [274] . Combining these two methods, PSIS and LOO-CV, yields a validation method called PSIS-LOO-CV [274], which is beneficial in situations like this, where an MCMC sampling-based model is trained on a small dataset. As a result, we get for each model the ELPD score, which we use to compare the three proposed models, a parameter η, which can be interpreted as the effective number of degrees of freedom in the model, and the so-called Pareto shape parameters k_i, which assess for each data point i in the dataset, how much it affects the ELPD estimation. For data points where k_i exceeds 0.7, the PSIS-LOO-CV estimate becomes unreliable, which can also indicate an under-constrained model or an outlier in the data [274].

8.3.5 Evaluation of feature importance

To estimate the importance that our model assigns to each feature of the reflectance spectra, we calculate a model-agnostic measure of feature importance [89] called *model reliance* (MR). Here, the importance of an individual feature is calculated as the relative change in the model's error when the individual observations of only that feature are shuffled, compared to the error on non-shuffled data. This *causal intervention* intentionally breaks the dependence between different correlated features. Therefore, MR, unlike correlation analysis, is a causal tool to diagnose the model rather than the data. This is relevant here because the different features of our model are computed by taking the inner product between the reflectance spectra and a set of overlapping not independent basis functions and are hence certainly correlated. We use the same loss function as for the model selection, namely ELPD. Since this measure already estimates the logarithm of a quantity of interest (the posterior density), we use the difference between shuffled and non-shuffled ELPD instead of their ratio to estimate the logarithm of the MR for the posterior density. Because we are only interested in qualitatively ranking features by their importance, we normalize the resulting importance value of each feature by their average. To improve the robustness of this measure, the shuffling is repeated multiple times (here, ten times), and the results are averaged. Repeating this procedure for each feature of a model yields positive scores for ranking all features by their importance.

8.4 Results

In this section, we evaluate each of the three models presented above, namely the non-hierarchical model, the model with hierarchical bias term, and the full hierarchical model.

8.4.1 Model predictions

First, we visualize the models' accuracy and ability to generalize in a model-agnostic way by directly plotting predictions against the corresponding measured "ground-truth" values. For this purpose, we randomly select 80 % of all available data (the training set, shown in blue) to infer model parameters, which we then use to predict the LAI for the remaining 20 % of the data (the test set, shown in green). Due to the probabilistic nature of our models, a full posterior predictive distribution is given for each data point, which we summarize in figure 8.5 **(A),(C), and (E)**. We can observe that all three models make reasonable predictions, i.e., that the predicted LAI grows in proportion to the measured LAI. Because all our generalized linear models assume that the *logarithms* of the LAI scores are homoscedastic, the standard deviation of the predictive distribution increases with the measured LAI, as well. Rather than the raw residuals $r_i^j = \hat{Y}_i^j - Y_i^j$, we therefore compute the relative residuals $\tilde{r}_i^j = {r_i^j}/{Y_i^j}$, each normalized by the corresponding measured LAI value Y_i^j, and summarize them in the cumulative histograms shown in figure 8.5 **(B),(D) and (F)**. For all three models, the relative residuals are similar between training and test set, which indicates that they generalize well.

8.4.2 Model comparison

To quantify the generalization error of all three models more accurately, we use the PSIS-LOO-CV method on *all* available data to estimate the ELPD on novel data. This procedure yields several highly informative measures, which are summarized in table 8.2. To verify the convergence of the sampling procedure for each model, we compare the marginal posterior distribution of each parameter across multiple chains and find no discrepancies or divergences (see also Supplementary Material, Figure S2D, Figure S3D, and Figure S4D). We can see that the highest ELPD (indicating the lowest generalization error) is achieved for the two hierarchical models, with little difference between them (-157.8 and -157.0, respectively, with a standard deviation of ≈ 11.5 each). Suppose, for the sake of argument, that for a similar dataset, we would select models based purely on the ELPD. In that case, it might be a matter of chance to pick the model with only a hierarchical bias term (as in this case) or the full hierarchical model.

However, due to the limited amount of training data and considering that the number of parameters ranges from 13 for the non-hierarchical baseline model to 17 for the model with hierarchical bias term to 61 for the full hierarchical model, we are also concerned with model complexity and the risk of overfitting. Since LOO-CV estimates generalization error directly, it does not need to explicitly penalize a large number of parameters, which is a significant advantage when

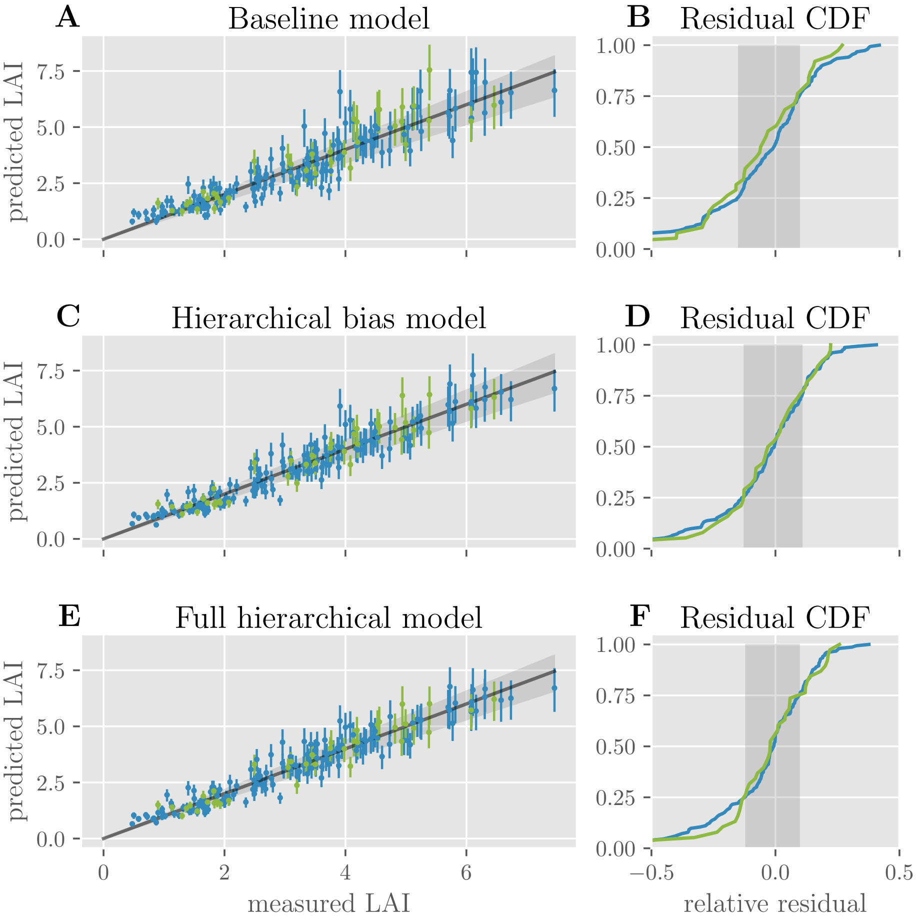


Figure 8.5. Model predictions of LAI. For the three models, **(A), (C) and (E)** plot for each data point (training data in blue and testing data in green) the predicted LAI values against the measured LAI values. Error bars indicate the interquartile range of the predictive distributions. Dots represent the expected value. The gray line shows the optimal predictions; the best 50% of model predictions lie within the gray cone around it. **(B), (D), and (F)** show the cumulative distribution function of the residuals, each normalized by the corresponding measured LAI value for training and testing data (blue and green lines). The gray areas show the same interquartile range as the cones in **(A), (C), and (E)**.

				Pareto k			
model	ELPD	#params.	η	(-Inf, 0.5]	(0.5, 0.7]	(0.7, 1]	(1, Inf)
Naive	-185.5±12.2	13	13.3	191	0	0	0
Hier. Full	-157.8±11.5	61	24.9	180	8	3	0
Hier. Bias	**-157.0±11.5**	**17**	**15.0**	**187**	**3**	**1**	**0**

Table 8.2. Comparison of the three models using PSIS-LOO-CV. The ELPD ± one standard deviation are listed for each model. #params denotes the number of parameters, η denotes the effective number of parameters. For each model, we show the number of data-points for which the Pareto shape parameter k falls into either of four different intervals.

comparing Bayesian hierarchical models. Instead, it allows us to estimate the model complexity of the three models by the so-called *effective number of parameters* η, which provides some intuition about how many degrees of freedom the model has to approximate the available data. As we see in table 8.2, $\eta = 13.3$ is quite close to the parameter count of the non-hierarchical model on the pooled dataset. This only increases slightly to $\eta = 15.0$ for the hierarchical bias model, even though it has four additional parameters. However, adding another 44 parameters for the full hierarchical model increases η substantially to 24.9.

Since PSIS-LOO-CV emulates conventional LOO-CV, it provides additional information that can help us understand how prone each model is to overfitting: For each data point, the procedure yields the shape parameter k of a Pareto distribution, which indicates whether estimating the generalization error for that data point is reliable ($k \in [-\infty, 0.7]$, ideally $k \leq 0.5$), potentially unreliable ($k \in [0.7, 1]$) or entirely unreliable ($k \in [1, \infty]$).[274] As table 8.2 shows, the full hierarchical model struggles with PSIS-LOO-CV for three data points, which may indicate that the model is more prone to overfitting to these potential "outliers" (see also Supplementary Material, Figure S5D).

As these numbers suggest, the model with a hierarchical bias term is the best choice because it is barely more complex than the non-hierarchical model, yet it performs at least as well as the full hierarchical model.

8.4.3 An interpretable kernel function

As outlined above, all three models derive their predictions of LAI from a weighted linear combination of features, which we compute by taking inner products between the measured reflectance spectra and a set of B-spline basis functions. These linear operations can be equivalently expressed as taking the inner product between each reflectance spectrum and an inferred *kernel function* $\kappa^j(\lambda)$, which provides a different, more interpretable perspective on the model.

To motivate this equivalent perspective, we look at how the reflectance spectra affect the linear predictors μ_i^j of the respective GLMs in equations 8.2 to 8.4, ignoring the contribution of the inferred bias terms here. For all three models[2], we

[2] To simplify notation, we write w^j and b^j for the (possibly) dataset-specific weight and bias terms, and set $w^j = w$ or $b^j = b$ for models that don't make these dataset-specific distinctions.

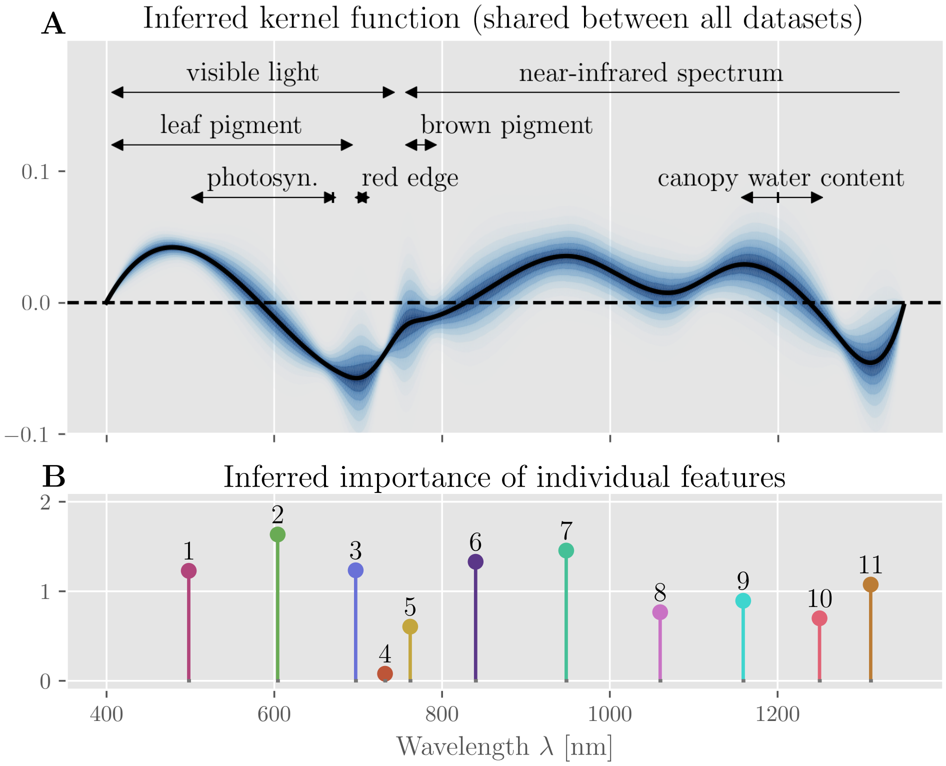


Figure 8.6. Inferred kernel function and feature importance. (A) shows the posterior distribution of the inferred kernel function. The black line represents the expected kernel. We can relate several ranges of the reflectance spectrum to physical phenomena, namely effects due to green leaf pigment (400 nm to 700 nm [275, 276]) and photosynthetic capacity (495 nm to 680 nm, peak at 670 nm [277, 275, 276]) and the red edge region (690 nm to 720 nm [278]) in the visible light range, as well as the canopy's water content (1150 nm to 1260 nm [279], peak absorption at around 1200 nm [277, 280]) in the near-infrared range. **(B)** shows a stem-plot of the relative importance of each feature (enumerated; normalized by the average feature importance) as well as the resulting estimated importance of each wavelength.

can use equation 8.1 to rewrite the contribution of the features extracted from the i^{th} reflectance spectrum R_i^j of the j^{th} dataset as follows:

$$\begin{aligned}\mu_i^j - b^j &= \sum_k (X^j)_{i,k} w_k^j \\ &= \sum_k \langle R_i^j, b_k \rangle w_k^j \\ &= \langle R_i^j, \kappa^j \rangle \quad \text{where } \kappa^j(\lambda) = \sum_k w_k^j b_k(\lambda)\end{aligned} \tag{8.5}$$

Since the parameters w_k^j are random variables, the kernel functions κ^j are random variables, samples of which can be generated by combining the (static) basis functions b_k with samples of w_k^j. Figure 8.6 (A) shows the distribution of the inferred kernel function for our model of choice, i.e., the hierarchical bias model (for the other two models, see Supplementary Material, Figure S6D, and Figure S7D). By analyzing this kernel function, we can identify regions of the reflectance spectrum that contribute positively or negatively (e.g., around $\lambda \approx 700\,\text{nm}$ and $\lambda \approx 1300\,\text{nm}$) to the predicted LAI score, and relate them to physical mechanisms.

8.4.4 Feature importance

In addition to the sign and magnitude of each feature's contribution (which are determined by the inferred weights; c.f. Supplementary Material, Figure S2D to Figure S4D), we are also interested in how important each individual feature is for the model's prediction. We quantify this via MR as described above. Figure 8.6 (B) shows that, except feature four, all features are indeed important for the prediction accuracy of the model. The low importance of feature four centered around 730 nm is likely due to the narrow domain of basis function b_4 (see figure 8.1 (D)), which indicates that this feature could be removed or an alternative knot-placement procedure could be chosen to reduce model complexity.

8.5 Discussion

Our results confirm that using a Bayesian hierarchical model not only leads to an improvement in the prediction accuracy over a non-hierarchical approach, but more importantly, it yields several qualitative benefits regarding interpretability, model complexity, and robustness.

One important benefit of the Bayesian hierarchical approach is that an appropriate choice of priors and model structure allows us to integrate additional model parameters without excessively increasing model complexity. For example, the number of spectral features used in our model directly determines the scale of the respective spline basis functions, which determines the resolution of our kernel function. This can create a trade-off between a model with lower spectral resolution and a model with a larger number of parameters. In the Bayesian approach, we can choose the model with more parameters without the risk of overfitting if we formalize our uncertainty and prior assumptions about the parameters appropriately. This is particularly important for hierarchical models, where we might want to add a large number of parameters to account for the specific variations in each subset of the data. Compare, for example, our full hierarchical model with its 61 parameters to the non-hierarchical baseline model with 13 parameters. Here, the addition of 48 new parameters only increases the effective degrees of freedom of the model by 11 and appears to increase the risk of overfitting only moderately. In our hierarchical bias model, we directly incorporate the fact that each of the four data subsets was recorded at a different growth stage of the plants, which affects the expected LAI, and hence requires a separate bias parameter. However, by simultaneously inferring the shared prior distribution over these separate parameters, we can ensure that the prediction on any one subset of the data benefits from the information contained in all the others. Of course, a non-hierarchical model can also benefit from heterogeneous data (see e.g. [262]), but it may fail in subtle ways if systematic differences *between* the data subsets obscure the relevant associations *within* each dataset. [3] In general, Bayesian hierarchical models allow us to conveniently include additional information about the dataset, domain knowledge, and regularizing priors, all of which can help to reduce the model's effective degrees of freedom. For the often small and heterogeneous datasets used in environmental sciences [231], this can be a major advantage over alternative machine learning approaches such as Random Forest Regression [282, 283] or Deep Learning [284, 285, 286], which may require prohibitive amounts of training data due to their typically large number of parameters.

We employ MCMC sampling to generate unbiased samples of the full posterior distributions over parameters and predictions, which allow us to use additional diagnostic tools and error measures. For example, we can directly estimate posterior densities, credible intervals, and even generalization errors via sample-based methods such as PSIS-LOO-CV, which are more broadly applicable than information criteria such as the Akaike Information Criterion (AIC), the Widely Applicable Information Criterion (WAIC), or the Bayesian Information Criterion (BIC) [287]. In particular, we saw that a hierarchical model might have a considerably larger number of parameters with a comparatively minor increase in model complexity, making any form of regularization based directly on the number of parameters difficult. Besides better diagnostics, sample-based measures can also provide insights about the data itself, e.g., indicating which data points are potential “outliers” that the model is susceptible to (see Supplementary Material, Figure S5D).

In addition to descriptive statistics, we also estimate feature importance using an intervention-based model-agnostic method that artificially breaks the dependence between naturally correlated features. Thus, we can infer exactly which features the model relies on for its prediction – independently from the

[3] In Supplementary Material, Figure S6D, we show that this is indeed the case here, as wavelength around 1200 nm lead to a pronounced dip in the kernel function when data from multiple datasets is pooled, but this association disappears if the model is instead fit to any individual dataset. Looking at the pooled dataset, we would therefore be led to conclude that lower spectral power around 1200 nm is a strong predictor of higher LAI. While this is correct on the artificially pooled dataset, it appears to be incorrect on any individual datasets. This may be an instance of Simpson's paradox [281], which suggests that a hierarchical model is more appropriate here.

magnitude of the respective parameters. Such information can help domain experts identify potential problems, e.g. if supposedly relevant features are ignored, or irrelevant features are relied on. This simple example shows how methods from causal analysis [288] can help explain or interpret the model in qualitatively different ways than descriptive statistics alone.

Because we use a generalized linear model, we can additionally analyze and interpret the model's linear predictor directly in the measurement space. Since the individual features are extracted from the spectra using B-spline basis functions, this linear predictor is just an inner product between a reflectance spectrum and a kernel function plus an additive bias term. Due to the logarithmic link function, the bias term ultimately has a scaling effect on the LAI predictions. The kernel function directly shows which wavelengths are associated with higher LAI, e.g., short wavelengths of the visible spectrum and much of the near-infrared spectrum, or lower LAI, e.g., around 600 nm to 750 nm. These results can be directly linked to physical phenomena and examined with domain knowledge. For example, the positive association for short wavelengths in the range 400 nm to 550 nm may be attributable to the effect of green leaf pigment, which reflects light in the range 400 nm to 700 nm [275, 276]. Similarly, the pronounced drop around the red edge (690 nm to 720 nm [278]), which is related to the plants' chlorophyll content [289, 290], total nitrogen [291, 292] and yield [293, 294], may be attributed to the plants' photosynthetic capacity (495 nm to 680 nm [277, 275, 276]) that peaks at around 670 nm.

Finally, we opted for a simplistic, interpretable model of LAI as a function of spectral power, but the hierarchical Bayesian modeling approach makes it easy to extend the proposed model further. For a larger dataset, the model complexity could be increased, either by choosing broader priors or by increasing the number of parameters to improve its accuracy, or the additional measurements could be used to reduce the uncertainty over the model's parameters. The data used in this study consists of measurements taken at four distinct locations and points in time. By allowing the specific parameters for each of these datasets to vary independently around a shared set of global parameters, we thus indirectly account for the combined effect of location and time. If a much larger number of datasets is used, their spatio-temporal distribution could also be taken into account explicitly, for example, by forcing the specific parameters of the datasets to be correlated, depending on their proximity in time and space, through an appropriate choice of a joint prior distribution. This approach leads to a full spatio-temporal model, which could also predict LAI on crop sites for which no training data is available. One could also include additional levels of hierarchy (e.g., to extend the model to other related plant species or different geographical regions), or other factors such as soil moisture content [295], the influence of climate change and CO2 concentration on crop growth [246, 247], the effects of warming asymmetry due to climate change [296], effects of microclimate [297], the influence of the amount of soil conditioner on the crops [298], and ammonium level in the soil.

Acknowledgments

We would like to thank our reviewers for their detailed and constructive feedback, which helped to improve this manuscript.

Chapter 9

Conclusion

In this thesis, I developed three prediction models for tasks from different scientific fields: prediction of epileptic seizures (medicine), prediction of infectious diseases (epidemiology), and predictions of the leaf area index (remote sensing / environmental sciences). Although they are from different fields, they face similar challenges, from data collection and understanding domain-specific context to developing interpretable machine learning models that yield good predictions. This thesis shows possible ways to tackle these challenges. To do so, I followed four principles that I think are invaluable when working on interdisciplinary data science projects.

1. Work on a challenging real-world problem with a real-world dataset

   Although benchmark datasets are very beneficial for developing machine learning models [28], the full potential of data science can only be explored when we go from benchmark to real-world datasets. Working on real-world datasets is more challenging, but it pays off in the long run because models developed on real-world datasets need to tackle more issues early on, e.g., how to build a model with limited or heterogeneous datasets. This helps because big curated datasets are not equally prevalent across research fields. [26]

2. Actively communicate and collaborate with domain experts

   In a collaboration between data scientists and domain experts, data scientists come from "the outside". Using their modern and interesting tools can help confront the challenges and phenomena of different scientific fields. But this only brings benefits when it is supported by domain expertise. Only domain experts understand the challenge thoroughly and can help others understand what the desired outcome of the model should be. Their help in interpreting data context and phenomena is crucial for model development and for understanding the implications of the model.

3. Incorporate data context diligently

   Data is situated in the environment in which it was created. [16, 15] To understand the data completely, one must know how and why it was collected and created and whether something was omitted. Incorporating data context is crucial because it helps to account for the specifics of the data and to avoid drawing unsupported conclusions.

4. Focus on interpretable models with components that can be visualized and understood

   In this thesis, I wanted to make models that would, besides offering accurate predictions, help us understand the underlying phenomena, which would also benefit domain experts. To do so, I focused on the interpretability of machine learning models. The interpretable components of the three models I developed can be visualized and interpreted by domain experts, which brings added value and distinguishes them from other models in the respective fields.

Here I want to briefly recount how I applied these principles in all three publications.

9.1 Predicting epileptic seizures

In the paper *"Predicting epileptic seizures using nonnegative matrix factorization"*, we showed the challenges of working with data from the clinical setting. Collecting medical data is, in general, time-consuming and takes a lot of resources, especially when recording epileptic seizures. Since epileptic seizures are rare events, even more time is needed to record them than usual, but the well-being of patients also limits the time frame. To then detect them, we need special handling of data since we are interested in events that would, in other settings, be seen as outliers. The task is further complicated since predictions have to be patient-specific, which limits the amount of data even further. One type of error is also more dangerous than the other (failing to warn about seizures is more dangerous than falsely predicting seizures).

To incorporate this challenging data context, we collaborated with one of the leading experts in seizure prediction, Dr. Levin Kuhlmann from Monash University in Melbourne. He helped us with applying our model to the Epilepsyecosystem dataset and added valuable domain expertise for interpreting preictal components. Our joint paper *"Predicting epileptic seizures using nonnegative matrix factorization"* has been, according to Google Scholar, cited 24 times by the end of January 2023.

To address the challenge, we started with the simple idea of using spectral analysis and decomposition to find distinctive features of preictal states, which we visualize and use as a defining feature for classification. We do this with nonnegative matrix factorization to learn the time and frequency components of each state. Although NMF learned interpretable differences in preictal and interictal states, this by itself wouldn't be enough to predict rare events. To further enhance our model, we used the SMOTE sampling technique to oversample preictal states during classification and to lessen the influence of interictal states.

Our paper shows that learning interpretable features of preictal and interictal states is possible. We specifically show how NMF lends itself to interpretable machine learning as a part of feature design. These results have a dual purpose: classification between preictal and interictal states, and learning about the dynamics of epileptic seizures. We tested our *locally interpretable* model on two real-word datasets: the EPILEPSIAE dataset [31] and the Epilepsyecosystem dataset [114], and showed that it yields good results on both.

9.2 Predicting infectious diseases

In the paper *"A Bayesian Monte Carlo approach for predicting the spread of infectious diseases"*, we showed how to approach data collection and data privacy challenges in epidemiology. Here we worked with real-world spatio-temporal data of three diseases. We showed how Bayesian models can be used to predict infections in time and space and learn the dynamics of various diseases. The paper demonstrates that it is possible to design transparent Bayesian models that are *globally interpretable on a modular level.*

During the project, we collaborated with the experts from the Signale Group of the Robert Koch Institute, Dr. Alexander Ullrich and Dr. Stéphane Ghozzi, who helped us with their domain expertise in epidemiology and working with real-world spatio-temporal data. As a part of the project, I have spent two months at the institute, where we developed and implemented the first version of the BSTIM model. There I had first-hand experience with how epidemiological data are monitored and collected, and I learned what domain experts are interested in learning from the model. Our joint paper *A Bayesian Monte Carlo approach for predicting the spread of infectious diseases* has been, according to Google Scholar, by the end of January 2023, cited more than 25 times. Further, I have presented a poster *Visualizing the spread of infectious diseases using public health data* at the EU Data Viz conference in 2019, organised by the Publications Office of the European Union. There I presented how to communicate epidemiological data and predictions of our model to the public, and the advantages such a model can bring.

As described in chapter *Example 2: predicting infectious diseases*, epidemiological data is at the intersection of spatio-temporal and public data, and developing machine learning models for such data relies heavily on data context. We addressed the data context in all steps of the process, from data preparation and analysis to the choice of our basis functions.

Spatio-temporal data are also quite different from other data types common in machine learning, such as images. We showed that probabilistic modeling and Bayesian statistics are good choices for spatio-temporal data since we can incorporate prior information in both dimensions. Our basis functions learn interactions in space and over time, as well as trends and seasonality of all three modeled diseases. Similar to my prior work on predicting epileptic seizures, our model has two outcomes: spatio-temporal kernels that we can visualize show us the disease dynamics, and we get predictions for each county and week. The interpretability that spatio-temporal kernels bring is the main strength of our model and distinguishes it from similar models in the field. We showed that it is possible to create an interpretable model that yields accurate predictions on par with a commonly used state-of-the-art model *hhh4*.

9.3 Predicting environmental variables

In the third paper, *“Bayesian hierarchical models can infer interpretable predictions of leaf area index from heterogeneous datasets”*, we applied a hierarchical Bayesian model to learn spatio-temporal differences within real-world heterogenous datasets in remote sensing. We first show how to use spline-based modeling of reflectance spectra to construct interpretable features and combine them in Bayesian hierarchical models to connect physical phenomena to LAI predictions. The paper demonstrates that it is possible to design a transparent and *globally interpretable model on a modular level* for accurate LAI predictions.

In this project, we collaborated with Dr. Bastian Siegmann from Jülich Research Centre and Dr. Thomas Jarmer from the Institute of Computer Science in Osnabrück, which provided us with data and gave valuable domain expertise. This was the first project where I worked on environmental variables, and having domain experts helped us conceptualize what interpretability could bring when inferring LAI from the reflectance spectra.

Data in this example consists of simultaneous measurements of reflectance spectra of white winter wheat and corresponding LAI. The measurements were carried out in four different years in two different fields. The data has a spatial and temporal component. However, it is not spatio-temporal data in the same sense as in the previous example of predicting infectious diseases. The epidemiological data we used for the BSTIM model is a time series of weekly infection counts. Here, data consists of discrete points in time, measured over a couple of days, in different fields, and multiple years, making data heterogeneous and limited.

Similar to the prediction of epileptic seizures, we started by designing simple and interpretable features, the modeled reflectance spectra. The splines with adaptive knots find points with “the most structure” in spectra, which we then use in a Bayesian hierarchical model. We show that Bayesian hierarchical models are well suited for incorporating the data context of limited and heterogenous datasets because, besides incorporating domain knowledge, they can also learn specific differences between datasets. This property is especially valuable in environmental sciences, where measurements are often taken from different locations and monitored over time. As in previous examples, our model has two outcomes, an interpretable kernel function of reflectance spectra, which we visualize, and LAI predictions. The kernel functions show regions of spectra that contribute positively or negatively to LAI predictions, which can be related to physical phenomena and understood by domain experts.

9.4 Broader conclusions from this thesis

Besides the impact of each paper in the respective field, I wanted this thesis to tackle broader issues, such as showing the importance and the advantages of collaborating with domain experts, recognizing the value of data context, and showing the potential of interpretable machine learning. While working on the projects, I came to more general conclusions, which can be summarized as:

1. Collaboration takes effort

   In all of the projects, I have collaborated with domain experts. I think the collaborations strengthen the models and the papers. However, collaboration on an interdisciplinary project can be challenging. Besides differences in scientific fields and their methods, there can be distrust of domain experts towards the new methods, especially black box models. [82] Using interpretable methods leads to more trust between the team members [81], which makes domain experts more likely to accept the proposed method.

2. Real-word data is "messy", but using an interpretable model can help

   In the projects, I worked with data types like time-series, spatio-temporal, public, and environmental data. We showed that the focus on interpretability most benefits the fields where data is limited or hard to collect. As previously described, real-world datasets are often heterogenous, limited, or contain meta information that would be hard for black box models to learn. By focusing on interpretability, we can incorporate the data context of such datasets, which significantly helps during the model development and for the interpretation of results.

3. Performance is not all that matters

   All of the models in this thesis have two outcomes: interpretable functions or components that give insight into the dynamics of the process, and predictions of the variable of interest. The experience of working with domain experts showed that they are interested in more than the high performing model. By focusing on interpretable models with components that can be visualized and understood, we help domain experts learn something new about the phenomena and add more value to the model.

4. The systematic approach enforced by interpretable machine learning allows us to make informed decisions about the model

   We can have more control during the model development by choosing an interpretable machine learning model in which we incorporate data context and domain expertise. This leads to a better understanding of what the model learns and how it uses features for predictions, which is crucial in troubleshooting and developing new versions of the model. Finally, it helps us understand and anticipate how predictions will be perceived, which is important for communicating results to the broader audience.

5. Data science is not a linear process

   The data science life cycle is an iterative, non-linear process. [71] Starting from identifying problems to data collection, processing, and analysis to developing and deploying a model, it can take a lot of iterations between the steps. Interpretable models make it easier to go back and forth between the steps, which makes the whole process faster and more transparent. [69]

9.5 Data science in the private sector

A noticeable difference between data science and other scientific fields is its multidisciplinary focus, because its methods are applicable to various domains. [28, 299] Even though the beginning of the field traces back as far as the 1960s, marked by the publication of *"The Future of Data Analysis"* by John Tukey [299], we can witness an increase in attention to data science and its possibilities following the rise of computing power and the interest of the private sector in predictive modeling. [99] The private sector brought further attention to data science by making large datasets publicly available, which became the foundation for benchmarking. [28]

This puts data science in a delicate position at the intersection of public and private interests. Data science is often directly influenced by the interests and needs of the private sector because it usually has more resources, especially companies large enough to have own research departments. Even though the private sector initially popularized black box models, I think interpretable machine learning can equally benefit the industry.

Since I have experience working in both the public and the private sector, I have noticed several differences between academia and industry that are worth mentioning here. The biggest difference comes from how the models are used and what happens with the predictions afterwards. Academia is concerned with researching new machine learning methods, their theoretical development, and testing new approaches. On the other hand, the industry focuses on producing and deploying well-established machine learning methods. A company or its clients further use the results to make business decisions.

It is possible to profit from the intellectual property of models and predictions, i.e., a company can sell machine learning models (like COMPAS, see chapter *What is interpretable machine learning*) or predictions (like weather predictions by BreezoMeter, see chapter *Why machine learning models (often) fail*). The difference is when a third party uses the model (in the case of COMPAS), there is no clear responsibility for wrong predictions. [33] A company that develops a model typically takes no responsibility for how a third party uses it or its outcomes. They also don't have the incentive to increase the model's transparency or interpretability, since disclosing the model's details would reveal trade secrets.

On the other hand, a company that sells predictions will often get more questions asking for explanations or justifications for trends in the data. This can be in the form of public outrage, as in the case of the UK grading algorithm (see chapter *Why machine learning models (often) fail*), or as business questions from a client that bought the data. In such situations, it is very useful to be able to go back to the model and see *why* it made specific predictions. Here, interpretable machine learning or a general focus on interpretability and transparency could prove valuable. Nevertheless, as long as there is no legal framework that requires companies to provide explanations in a standardized way, the share of responsibility for damages and harm caused by wrong predictions will remain unclear. [33] This is where I expect more pressure from the public or users.

9.6 The future of interpretable machine learning

9.6.1 The importance of AI regulations

Although there has been more research interest in interpretable machine learning in recent years, legal, technological, and ethical challenges in the field still require attention. The current lack of incentive to create interpretable models mainly comes from two factors: the intellectual property of black box models, and undefined responsibility for predictions of black box models. [69] As long as it is possible to profit from individual predictions and the intellectual property of proprietary black box models, there is no financial incentive to develop and offer interpretable machine learning models on the market. [33, 69] And as long as there is neither financial nor legal incentive for the transparency of machine learning algorithms on the market, there will be fewer interpretable models.

A way to change this is to enforce laws requiring algorithmic transparency for high stakes decisions. Currently, no country or entity enforces such laws. However, in 2018 the EU implemented the General Data Protection Regulation (GDPR), which forces companies, governments, and other entities to inform users of their data collection. [300, 301] Article 22 of GDPR states that "data subjects" (i.e., users) should not be subjected to automated decisions and that users, in general, have "a right to explanations." [302] However, it is not clearly stated that provided explanations must be accurate or what the explanations should look like. [303] The EU is currently working on a draft of a new, more detailed law that would require providers and users of AI systems in high stakes decisions to follow more strict rules on data, governance, transparency, human oversight, accuracy, and security. [304, 305] The law would explicitly prohibit AI systems that cause harm or exploit vulnerabilities of groups of people. [306] Further, the EU's AI Liability Directive, a law planned to be enforced in a couple of years, would make it possible for people to sue companies if they can prove that their AI harmed them.[1] [308]

The White House Office of Science and Technology Policy recently released The Blueprint for an AI Bill of Rights. [309] It is a white paper that describes five principles for the design, use, and deployment of automated systems. The principles say that AI systems should be removed if proven that they cause harm or are ineffective. Users should not face discrimination from algorithms, and they should be protected from "abusive data practices" and have agency over how their data are used. Further, users should also know when an automated system is used, understand how it contributes to the outcome, and it should be possible to opt out of AI systems where a human alternative would be better. [309] However, this is not (yet) legislation but just a set of guidelines for companies and entities. [310]

China has, since March 2022, enacted some regulations for explainable AI. [311] The law is known as Internet Information Service Algorithmic Recommendation Management Provisions, and it was drafted by the The Cyberspace Administration of China. [311] It requires that providers of recommendation services uphold users' rights. For example, the law prohibits using personal characteristics to offer different prices for a product. [312] The law further requires

[1] However, some non-governmental organizations worry that the hurdle to prove something like this would be too high for consumers, which would reduce the effectiveness of the law. [307]

that the applications do not promote content that encourages addictive behavior. This is a general requirement and could influence how social media works, but so far, there has not been much visible change. [313]

The public in the countries mentioned above also calls for more transparency in AI systems, which in turn influences governments to put more pressure on regulating AI. Researchers will have to focus more on interpretable machine learning because it seems likely that machine learning models will have to comply with upcoming and stricter laws in the future.

However, interpretable machine learning models are not fair by default, and building interpretable models might seem more complicated. [69] They also shift responsibility for the predictions to the model creators, but companies might be afraid of that.

On the other hand, interpretable machine learning can help design more ethical models, since they give more control to the model creators because they can see how the model works internally. Black box models have to deal with the same data complexities and challenges, but they usually show their shortcomings only *after* we already deployed them. [69] Because of this, the investment in interpretable machine learning pays off in the long term, and particularly in high stakes decisions.

9.6.2 The importance of multidisciplinarity

Another aspect that I think is important for the future development of the field is the possibility of applying similar data science tools to various tasks and collaborating with other scientific fields. This is also the aspect I wanted to show in this thesis by collaborating with domain experts from three scientific fields. The knowledge exchange should go both ways since also data science can benefit a lot from experiences in other fields. Here I will mention two concepts that I think will prove useful in the future: *datasheets for datasets* and *science about data science.*

First, starting from data and handling large datasets, AI researcher Timnit Gebru has suggested *"datasheets for datasets"*. [314] The concept is inspired by a common practice in electrical engineering: datasheets are instruction manuals for electronic components containing their operating characteristics, test results, recommended usage, and similar. [314] The idea is to provide a sheet with additional information for every dataset to ease the usage and provide the most accurate contextual information. Datasheets should contain information about the creation of a dataset, collection processes, and potential adjustments during data collection, e.g., what has been excluded from the dataset and why. Knowing the limitations of datasets is essential, and that contextual knowledge can help us build better models to analyze the data.

The authors argue that such datasheets would address the needs of data creators, data consumers, policy makers, investigative journalists, or other individuals. Data creators would have a standardized method of creating a dataset, which would include careful consideration of the process and the environment they want to collect data about, how to maintain the dataset, the implications of using the data, etc. Data consumers would have information about data context, which they can use for model development. Policy makers and other individuals (who don't necessarily have a background in data science) would have easier access to datasets. They could better understand their impact or the impact of machine learning models trained on these data. Datasheets for datasets would also help with the reproducibility of machine learning results, increase transparency, and would, in the long run, help with accountability and interpretability. Some researchers have already published datasheets with their datasets [315, 316, 317], and companies such as Microsoft, Google and IBM have internally created pilot versions of datasheets. [314]

Second, to standardize the field of data science, statistician David Donoho suggests the creation of a new scientific field he calls *"science about data science"*. [28] The field should focus on two main aspects: collection and curation of datasets and meta-analysis of methods. It should further introduce methods for meta-analysis, standardization, and rigorous evaluation of results. [28] He writes:

> *Data scientists are doing science about data science when they identify commonly occurring analysis/processing workflows, for example, using data about their frequency of occurrence in some scholarly or business domain; when they measure the effectiveness of standard workflows in terms of the human time, the computing resource, the analysis validity, or other performance metric, and when they uncover emergent phenomena in data analysis, for example, new patterns arising in data analysis workflows, or disturbing artifacts in published analysis results. The scope here also includes foundational work to make future such science possible - such as encoding documentation of individual analyses and conclusions in a standard digital format for future harvesting and meta-analysis. As data analysis and predictive modeling become an ever more widely distributed global enterprise, "science about data science" will grow dramatically in significance.*[28]

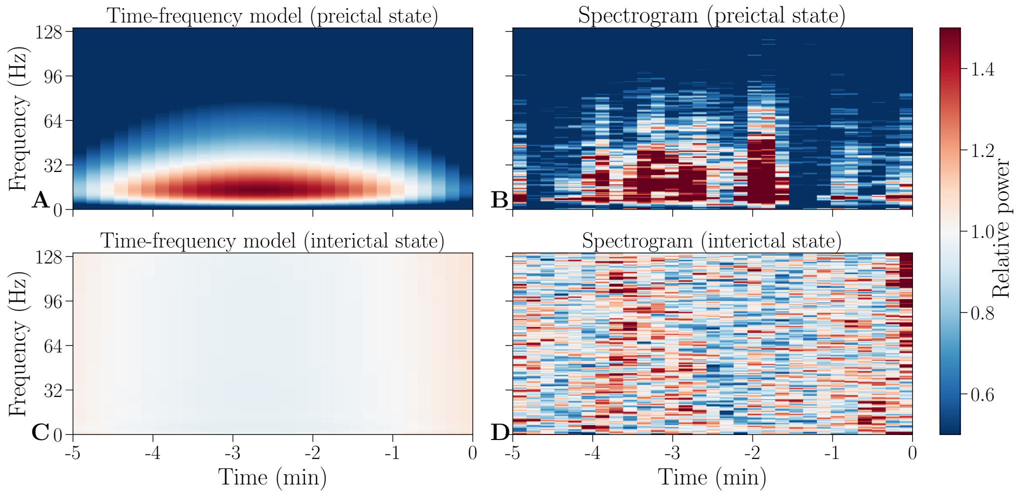
Science about data science would also help with reproducibility in machine learning and academia in general. Standardizing methods for data collection and analysis of workflows would also increase transparency and interpretability in machine learning. However, this is a great challenge which requires a lot of time and resources. A good example of data standardization is the Observational Medical Outcomes Partnership Common Data Model, which is an open community data standard for clinical data. [318]

Finally, I think the machine learning community has to work closely with researchers who study the impact of technology and automated decisions on our lives in the following years. Unfortunately, there has been a common sentiment in the machine learning community that it does not need input from other scientific fields to create high quality solutions. Geoffrey Hinton famously said in 2016: *"We should stop training radiologists now, it's just completely obvious within five years deep learning is going to do better than radiologists."*. [319] Six years later, the number of radiologists increased by 7% in the US, and (only) around 30% use AI as help in their work. [320] Similar was expected from AI during the coronavirus pandemic, yet most models are still unsuitable. [2, 3, 4, 5] I think this perceived exceptionalism is wrong: machine learning can not be successful in isolation. To create useful solutions and to use AI more extensively, we need to take the concerns of the public and government requirements seriously, work with domain experts, and develop more interpretable machine learning models.

Appendix A

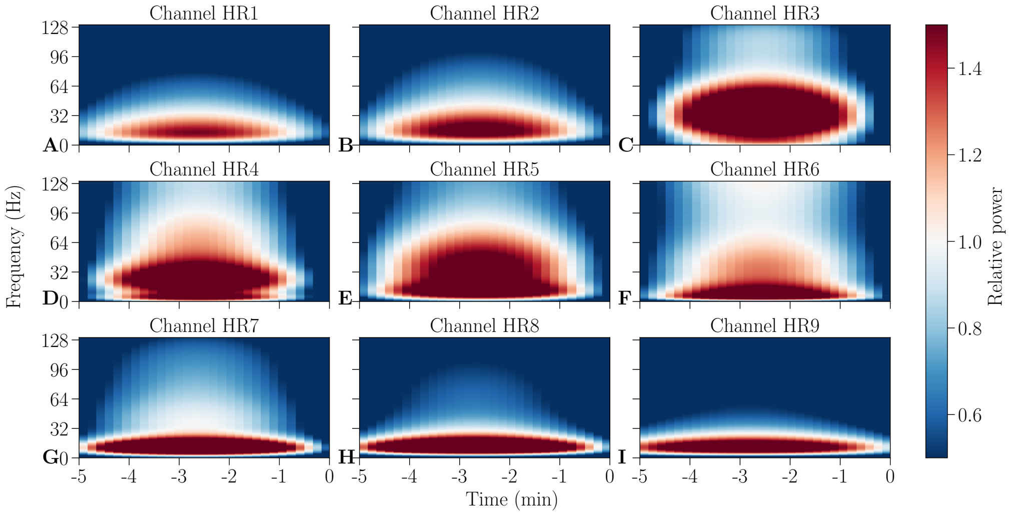
Supporting information: Predicting epileptic seizures using nonnegative matrix factorization

A.1 S1A Fig.



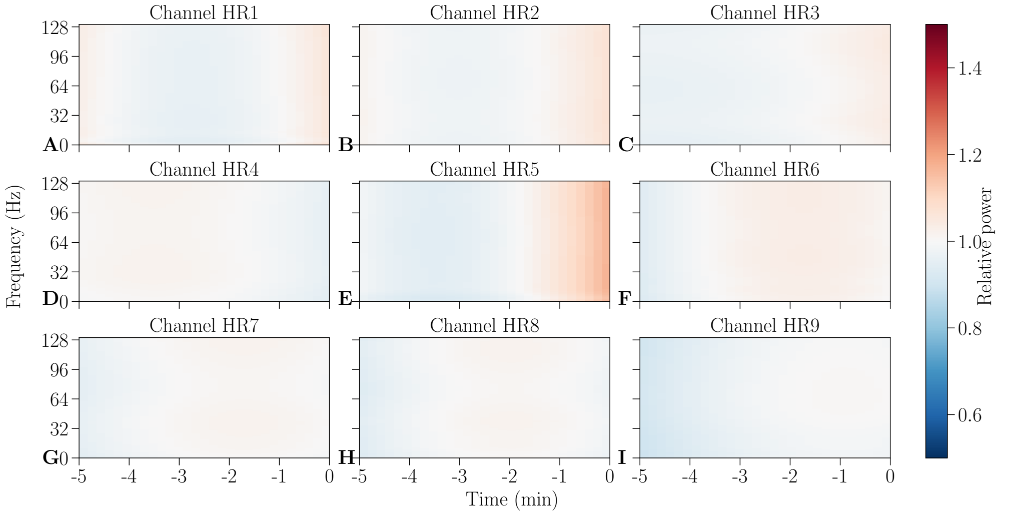
Time-frequency models and corresponding spectrograms of preictal and interictal states. An outer product of modeled time and frequency components (**A, C**) and corresponding spectrograms (**B, D**). A preictal state is shown in the upper row (**A-B**) and an interictal state is shown in the bottom row (**C-D**).

A.2 S2A Fig.



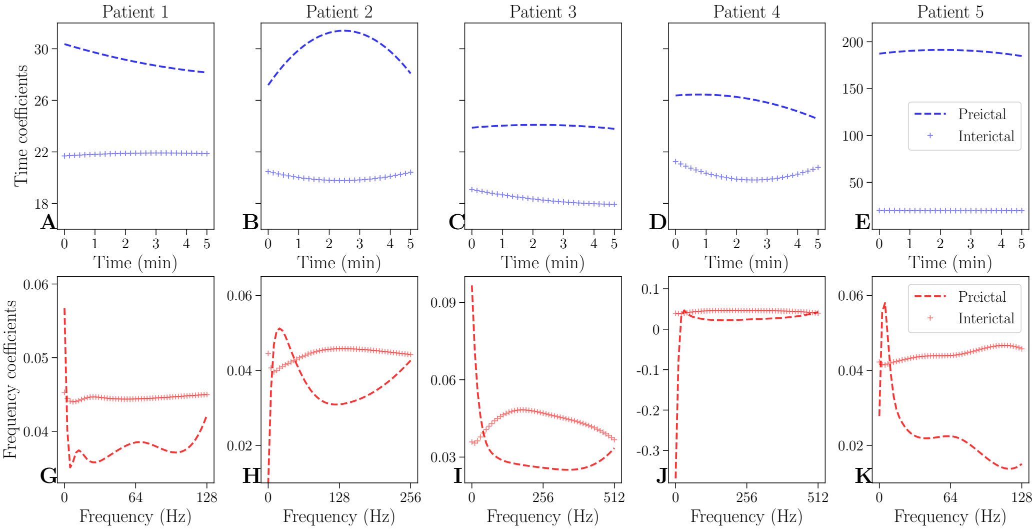
Models of preictal states. Models shown here are for different channels (**A-I**) from the same individual measurement period for patient 1.

A.3 S3A Fig.



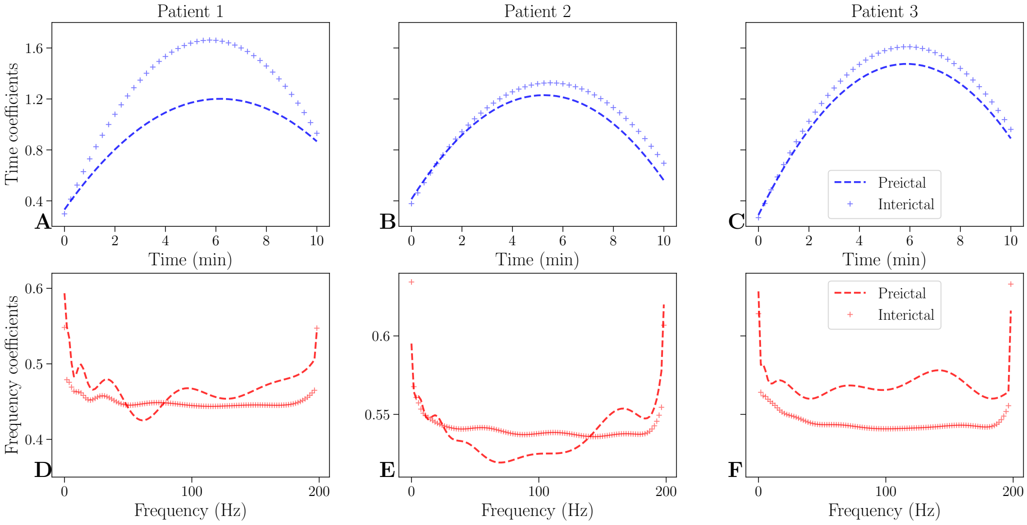
Models of interictal states Models shown here are for different channels (**A-I**) from the same individual measurement period for patient 1

A.4 S4A Fig.



Average models of time and frequency components of all channels and all measurements for preictal and interictal states of the EPILEPSIAE dataset. Models of time components are shown in the upper row (**A-E**), and models of frequency components are shown in the bottom row (**G-K**). Preictal states are indicated with a dashed line and interictal states are indicated with a line marked with + in blue for models of time and red for models of frequency components, respectively.

A.5 S5A Fig.



Average models of time and frequency components of all channels and all measurements for preictal and interictal states of Epilepsyecosystem dataset. Models of time components are shown in the upper row (**A-C**), and models of frequency components are shown in the bottom row (**D-F**). Preictal states are indicated with a dashed line and interictal states are indicated with a line marked with + in blue for models of time and red for models of frequency components, respectively.

Appendix B

Visualizing the spread of infectious diseases using public health data

UNIVERSITÄT OSNABRÜCK

Institute of Cognitive Science

Olivera Stojanovic ostojanovic@uos.de @stojanovic_olja

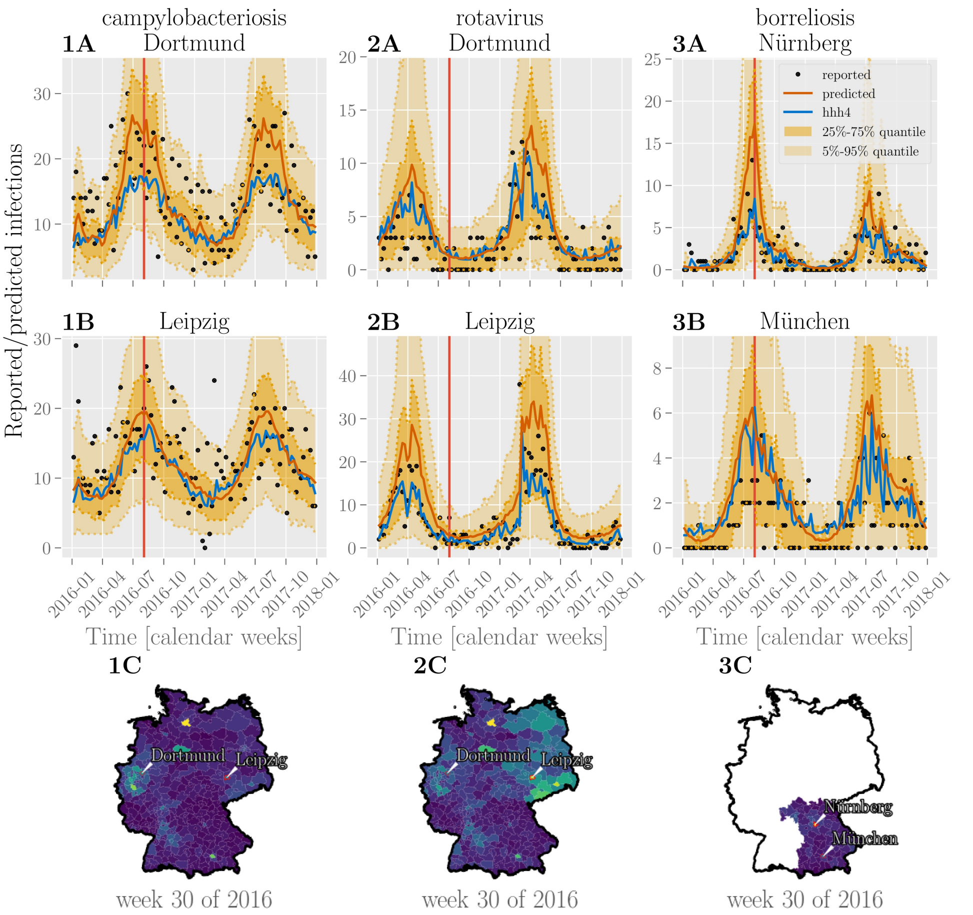
Visualizing the spread of infectious diseases using public health data

Public-health agencies have the responsibility to detect, prevent and control infections in the population. The Robert Koch Institute [1] in Germany collects a wide range of factors, such as location, age, gender, pathogen, and further specifics, of laboratory confirmed cases for approximately 80 infectious diseases through a mandatory surveillance system [2]. This data is publicly available, but in order to be useful for a broader public, it should be processed and presented in an interpretable form, using data visualizations and interactive tools.

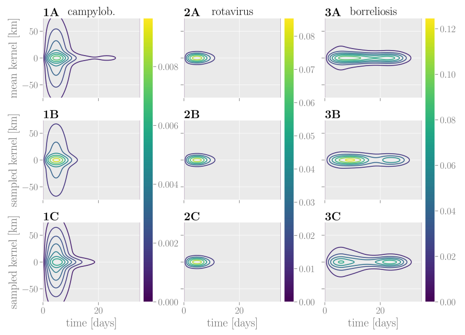
We develop a single model for predictions of infectious disease cases in all counties in Germany. To present the results of the model and to communicate potential risks of infection, **disease prediction maps** show a broad overview of the disease development in space, while plotting **prediction curves** together with collected data points show the disease development in time.

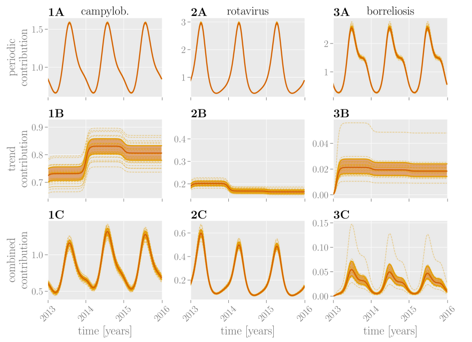
Bayesian spatio-temporal interaction model

The presented Bayesian spatio-temporal interaction model (BSTIM) [3] is a probabilistic model which predicts aggregated case counts within counties and calendar weeks. To this end, **publicly available health data**, region-specific and demographic data are used. We evaluated the BSTIM on a **one-week-ahead prediction** task for two diseases (campylobacteriosis [4] and rotaviral enteritis [5]) across Germany and for Lyme borreliosis [6] across the federal state of Bavaria. The BSTIM model predicts how many people are expected to become infected during the next week, in each county (shown in the upper figure). In addition, it provides **uncertainty estimates**, which give a sense of how confident the model is. Domain experts, in addition to model predictions, can access the dynamics and evolution of diseases. Visualizing learned model components (shown in the lower left figure) allows to inspect how diseases spread in time and space, while visualizing learned trends and seasonality (shown in the lower right figure) allows to see temporal evolution of diseases over the years. This **transparency and interpretability of machine learning models** increase scientific understanding and safety [7].



Predictions of case counts for various diseases by county. Reported infections **(black dots)**, predictions of case counts by BSTIM (orange line) and the hhh4 reference model (blue line) for campylobacteriosis **(column 1)**, rotavirus **(column 2)** and borreliosis **(column 3)** for two counties in Germany (for campylobacteriosis and rotavirus) or Bavaria (borreliosis), are shown in **rows A and B**. The shaded areas show the inner 25%-75% and 5%-95% percentile. **Row C** shows predictions of the respective disease for each county in Germany or the federal state of Bavaria in week 30 of 2016 (indicated by a vertical red line in **rows A and B**).





Key points:

- We develop and present a probabilistic model for prediction of infectious disease cases
- Prediction maps or risk awareness maps are beneficial for communicating a message to the public
- Visualizations of components of a prediction model are useful for domain experts for assessing the dynamics of diseases
- The machine learning system for prediction of infectious diseases should at the same time optimize for two aspects:
  - the prediction accuracy,
  - interpretability

References:

[1] https://www.rki.de/DE/Home/homepage_node.html

[2] Faensen D, Claus H, Benzler J, Ammon A, Pfoch T, Breuer T, et al (2006) **SurvNet@RKI – a multistate electronic reporting system for communicable diseases.** Euro surveillance: bulletin européen sur les maladies transmissibles=European communicable disease bulletin. 11(4):100–103.

[3] Stojanovic O, Leugering J, Pipa G, Ghozzi S, Ullrich A. (2019). **A Bayesian Monte Carlo approach for predicting the spread of infectious diseases**. Biorxiv; under review (PLOS ONE).

[4] https://www.who.int/news-room/fact-sheets/detail/campylobacter

[5] https://www.who.int/immunization/diseases/rotavirus/en/

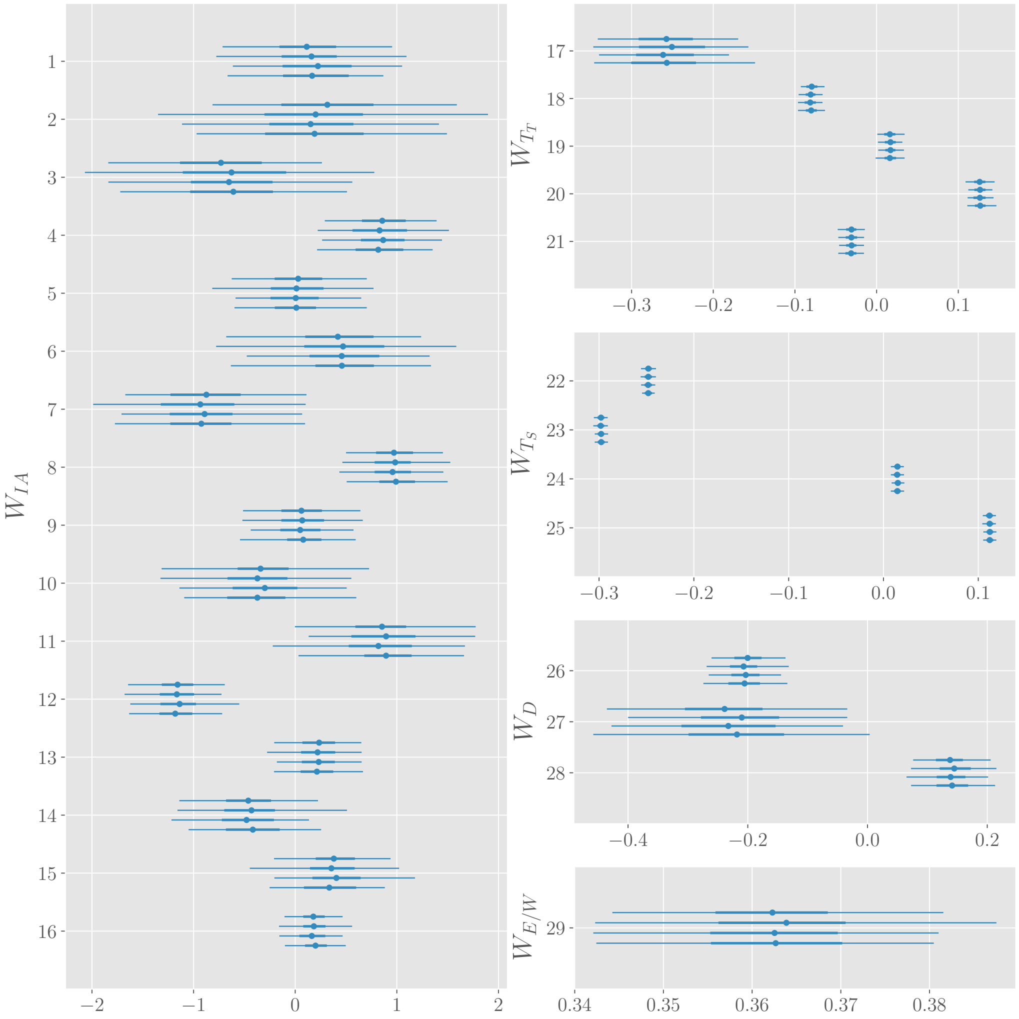
[6] https://www.who.int/ith/diseases/lyme/en/

[7] Molnar C. (2019) **Interpretable machine learning. A Guide for Making Black Box Models Explainable** (Chapter 2).

Appendix C

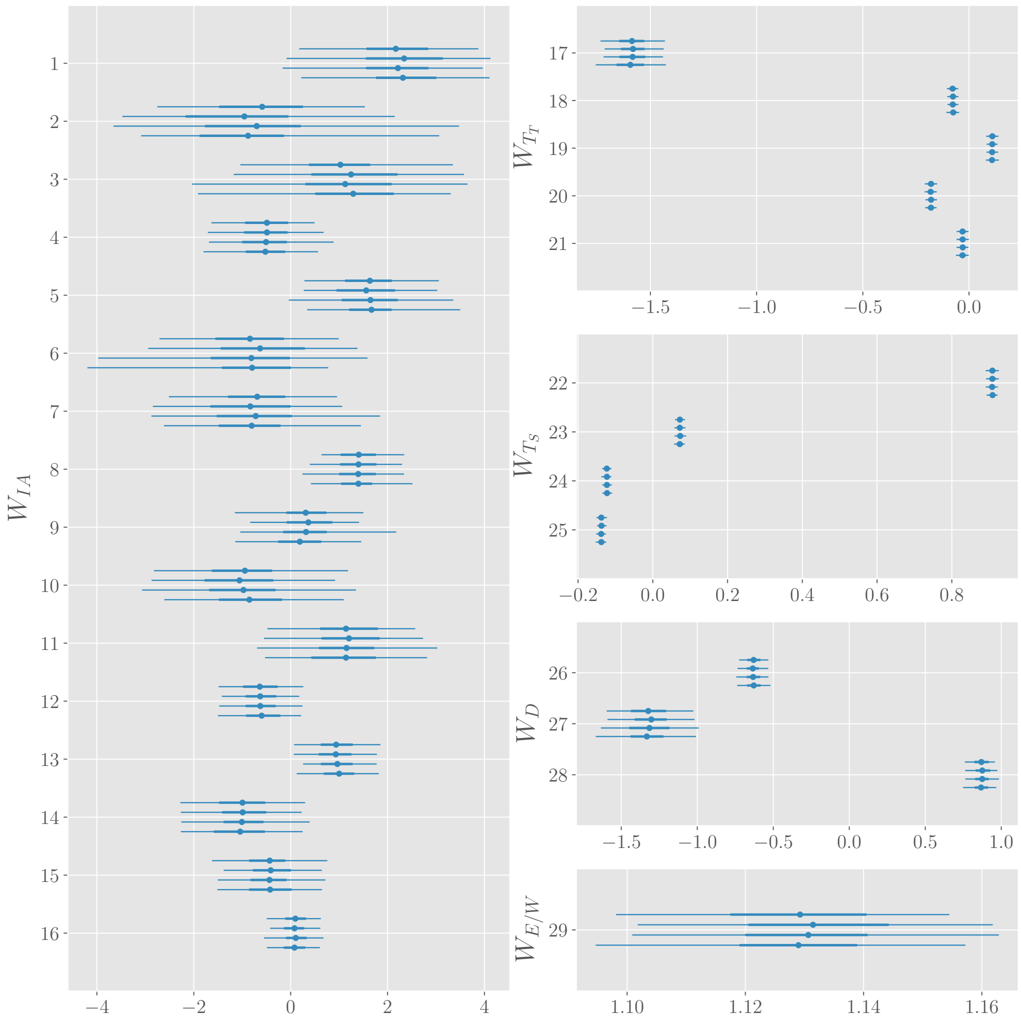
Supporting information: A Bayesian Monte Carlo approach for predicting the spread of infectious diseases

C.1 S1C Fig.



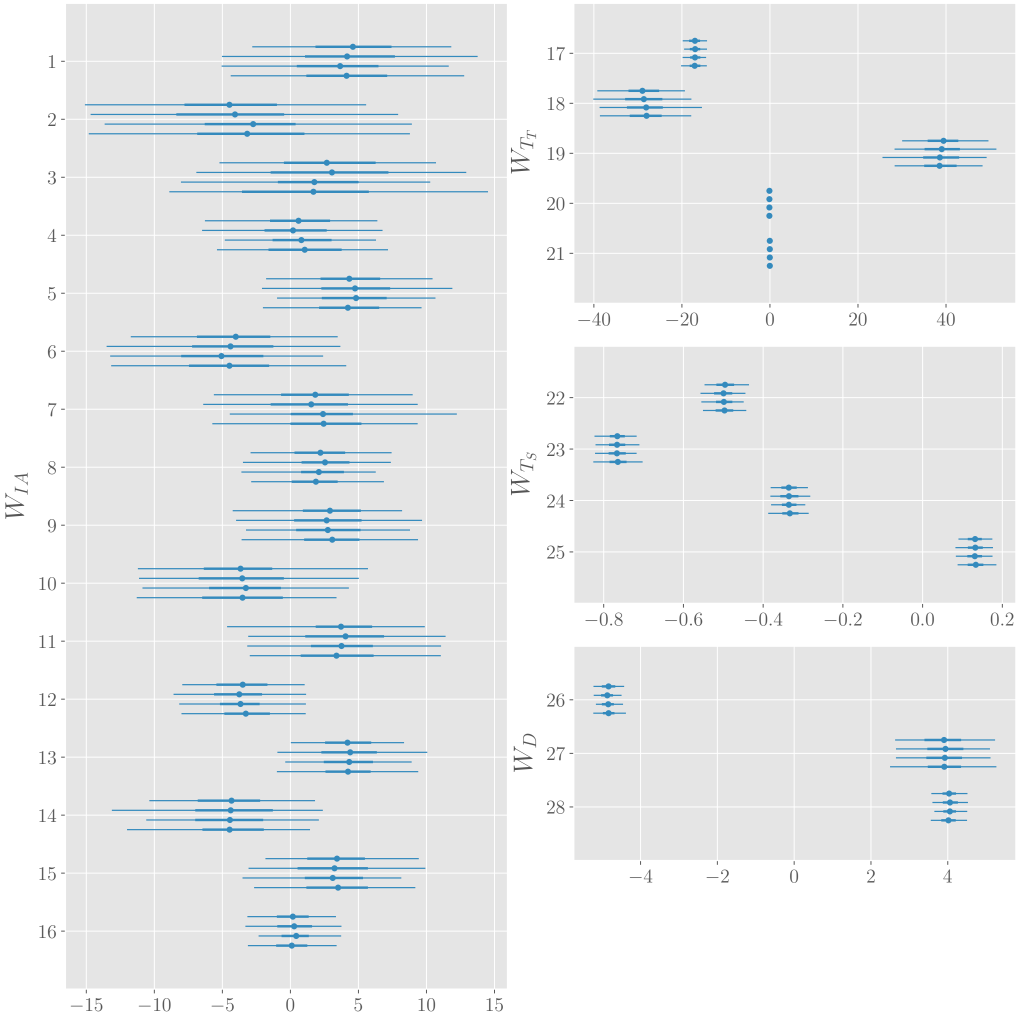
Marginal posterior distributions of all parameters for campylobacteriosis. For each of four Markov chains, the mean (dot), the range from the 25% to 75% percentile (thick horizontal lines) as well as the 2.5% to 97.5% percentile (thin horizontal lines) are shown. For all parameters, these summary statistics of the marginal distribution are similar across all four chains, indicating convergence of the MCMC sampling scheme (see also supplementary S7C Fig).

C.2 S2C Fig.



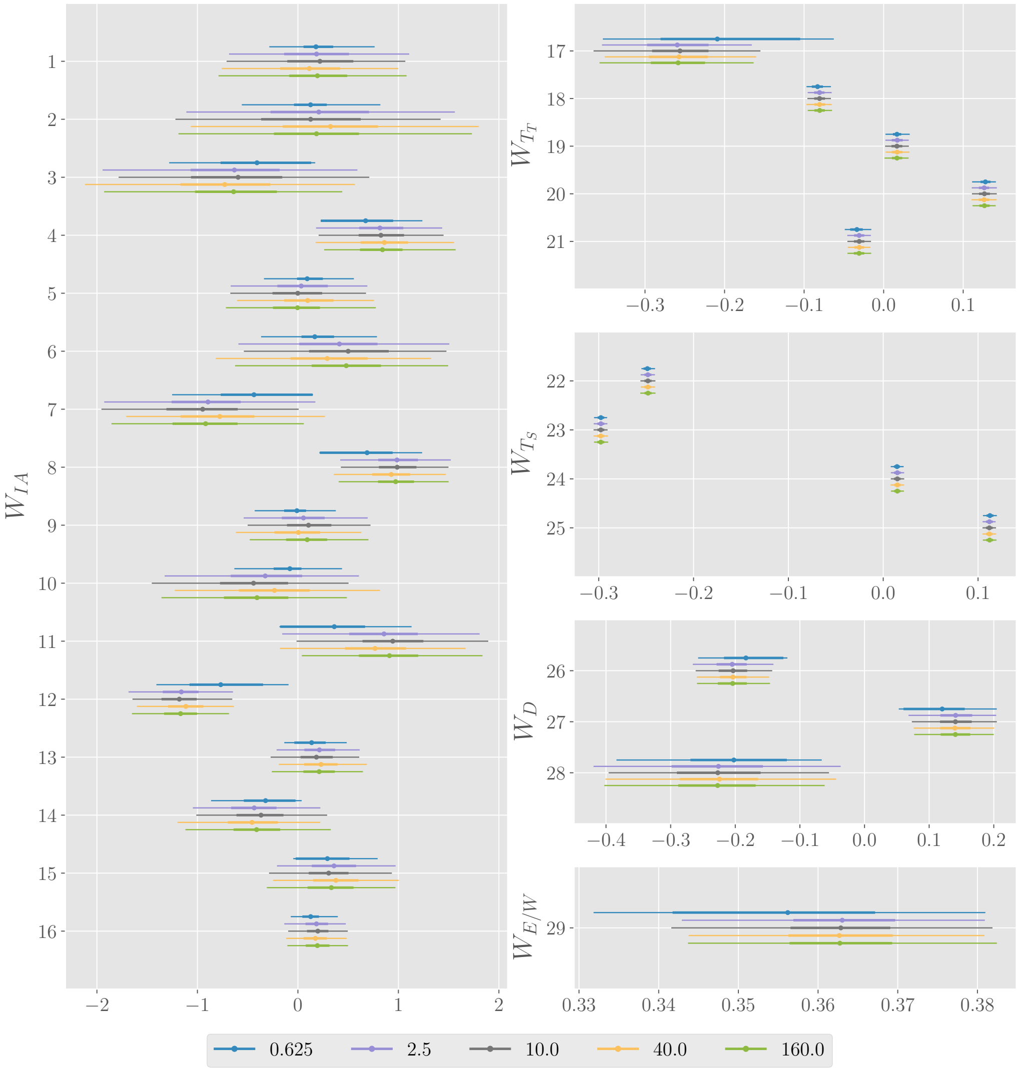
Marginal posterior distributions of all parameters for rotavirus. For each of four Markov chains, the mean (dot), the range from the 25% to 75% percentile (thick horizontal lines) as well as the 2.5% to 97.5% percentile (thin horizontal lines) are shown. For all parameters, these summary statistics of the marginal distribution are similar across all four chains, indicating convergence of the MCMC sampling scheme (see also supplementary S7C Fig).

C.3 S3C Fig.



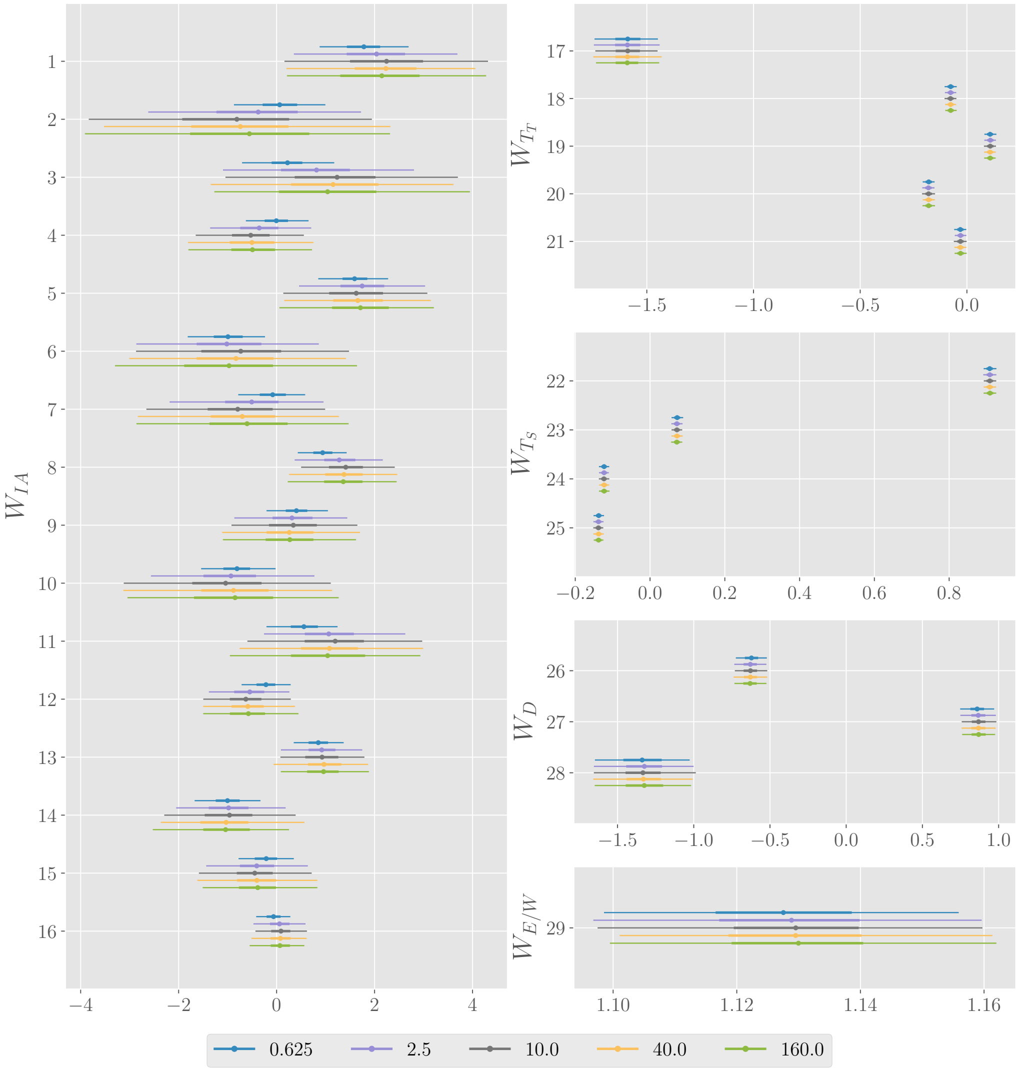
Marginal posterior distributions of all parameters for Lyme borreliosis. For each of four Markov chains, the mean (dot), the range from the 25% to 75% percentile (thick horizontal lines) as well as the 2.5% to 97.5% percentile (thin horizontal lines) are shown. For all parameters, these summary statistics of the marginal distribution are similar across all four chains, indicating convergence of the MCMC sampling scheme (see also supplementary S7C Fig).

C.4 S4C Fig.



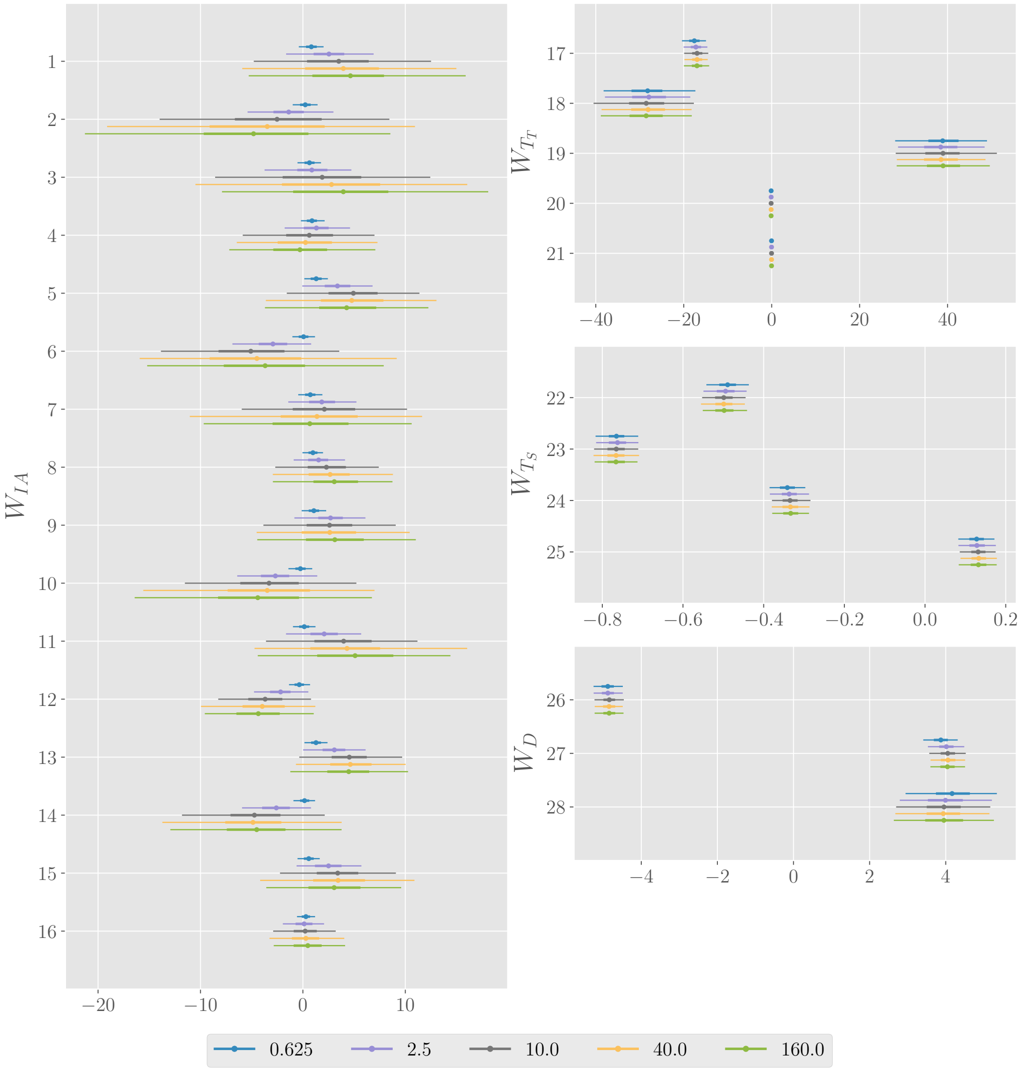
Sensitivity analysis for campylobacteriosis. Marginal posterior distributions of all parameters are shown for five different scales $\sigma_{W_{\text{IA}}} = \{0.625, 2.5, 10.0, 40.0, 160.0\}$ (color coded), which includes the special case $\sigma_{W_{\text{IA}}} = 10$ (see also supplementary S1C Fig) as used throughout this paper. For priors with standard deviation larger than 2.5, there is little qualitative change in the posterior distribution.

C.5 S5C Fig.



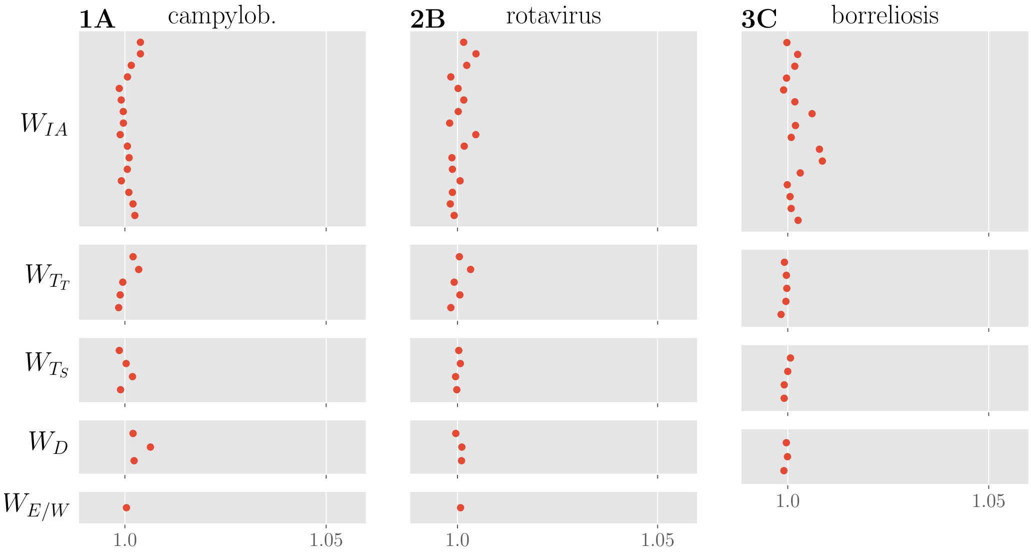
Sensitivity analysis for rotavirus. Marginal posterior distributions of all parameters are shown for five different scales $\sigma_{W_{\mathrm{IA}}} = \{0.625, 2.5, 10.0, 40.0, 160.0\}$ (color coded), which includes the special case $\sigma_{W_{\mathrm{IA}}} = 10$ (see also supplementary S2C Fig) as used throughout this paper. For priors with standard deviation larger than 2.5, there is little qualitative change in the posterior distribution.

C.6 S6C Fig.



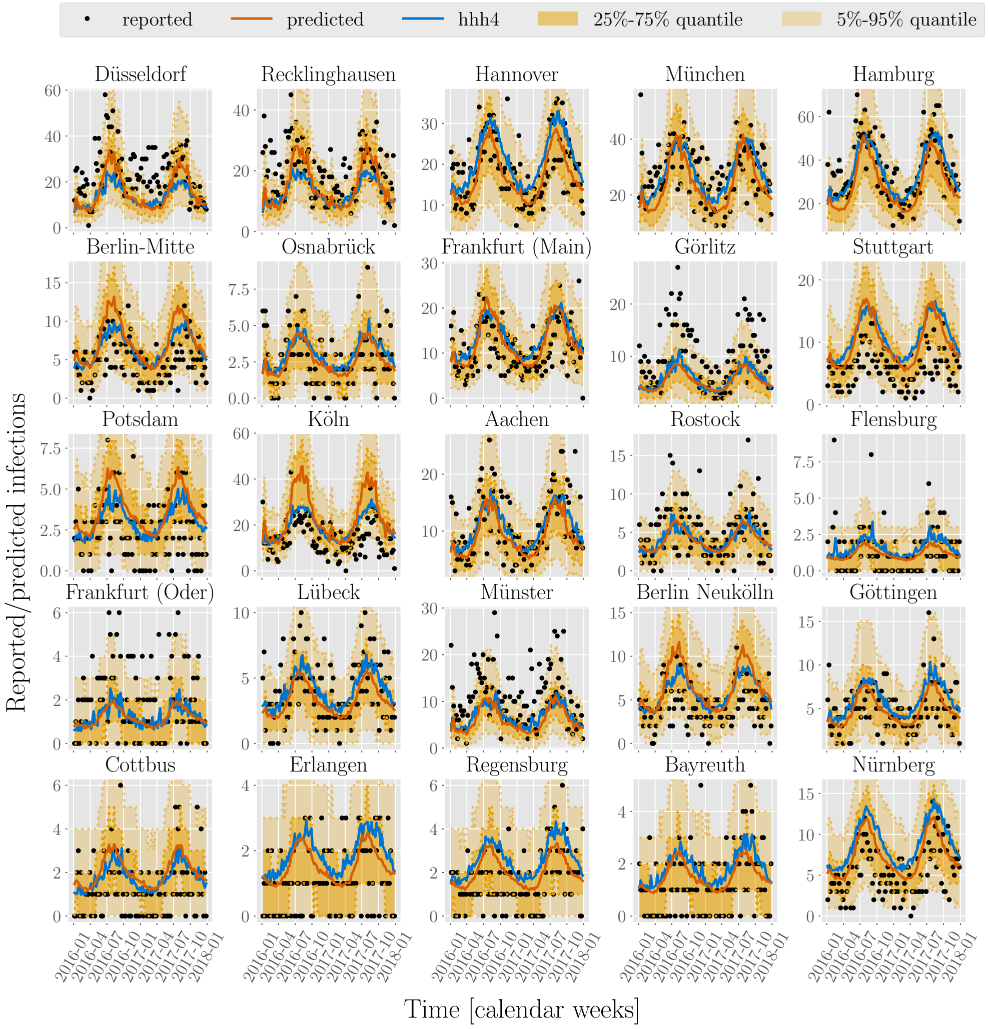
Sensitivity analysis for Lyme borreliosis. Marginal posterior distributions of all parameters are shown for five different scales $\sigma_{W_{\mathrm{IA}}} = \{0.625, 2.5, 10.0, 40.0, 160.0\}$ (color coded), which includes the special case $\sigma_{W_{\mathrm{IA}}} = 10$ (see also supplementary S3C Fig) as used throughout this paper. Here, the choice of prior has considerably more impact on the posterior distribution than for campylobacteriosis (see supplementary S4C Fig) or rotavirus (see supplementary S5C Fig), for both of which more training data is available. For a narrow prior with standard deviation 0.625, the interaction effect coefficients appear to be strongly regularized towards zero.

C.7 S7C Fig.



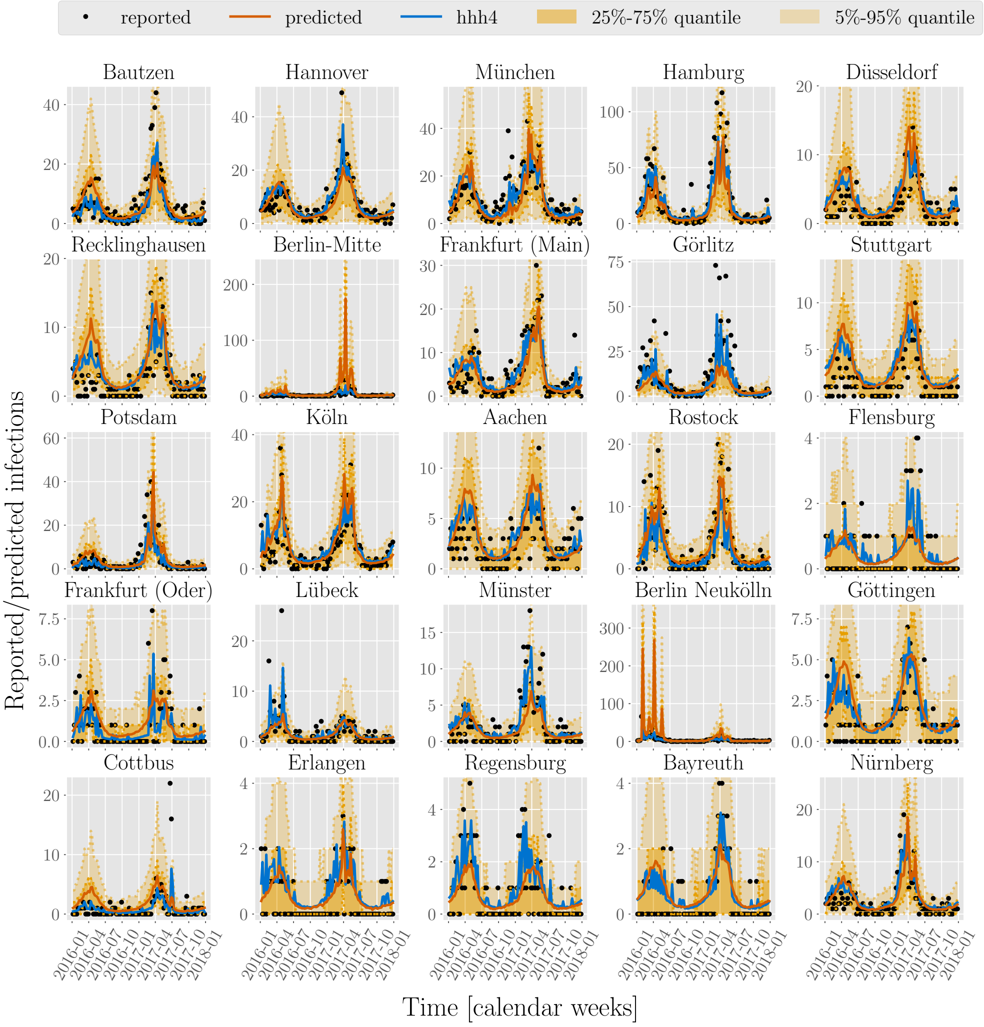
Convergence Diagnostics of MCMC chains. Gelman-Rubin diagnostics (red dots) for all parameters for campylobacteriosis (**1A**), rotavirus (**2B**) and borreliosis (**3C**). The values all lie close to 1.0 for all parameters, indicating convergence of the sampling procedure.

C.8 S8C Fig.



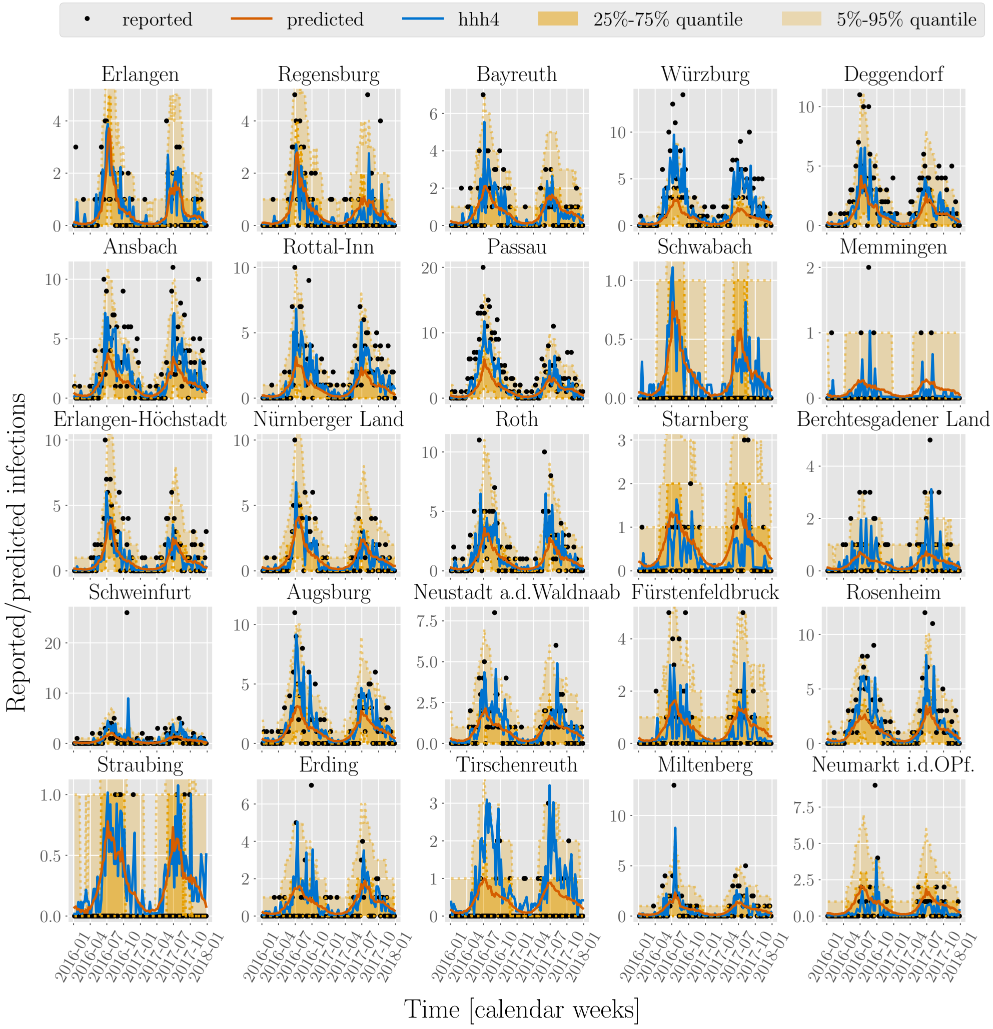
Predictions of case counts for campylobacteriosis for various counties across Germany. Reported infections (black dots), predictions of case counts by BSTIM (orange line) and the *hhh4* reference model (blue line) for campylobacteriosis for 25 counties in Germany. The shaded areas show the inner 25%-75% and 5%-95% percentile.

C.9 S9C Fig.



Predictions of case counts for rotavirus for various counties across Germany. Reported infections (black dots), predictions of case counts by BSTIM (orange line) and the *hhh4* reference model (blue line) for rotavirus for 25 counties in Germany. The shaded areas show the inner 25%-75% and 5%-95% percentile.

C.10 S10C Fig.

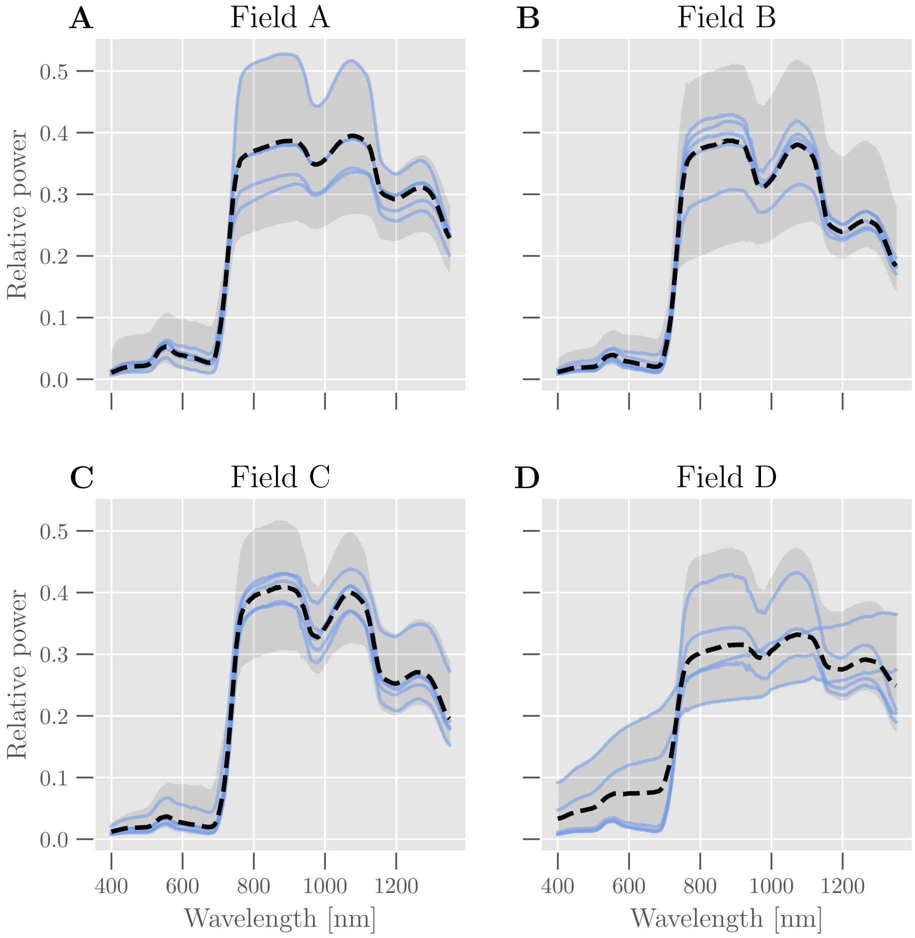


Predictions of case counts for borreliosis for various counties across Bavaria. Reported infections (black dots), predictions of case counts by BSTIM (orange line) and the *hhh4* reference model (blue line) for borreliosis for 25 counties in Bavaria. The shaded areas show the inner 25%-75% and 5%-95% percentile.

Appendix D

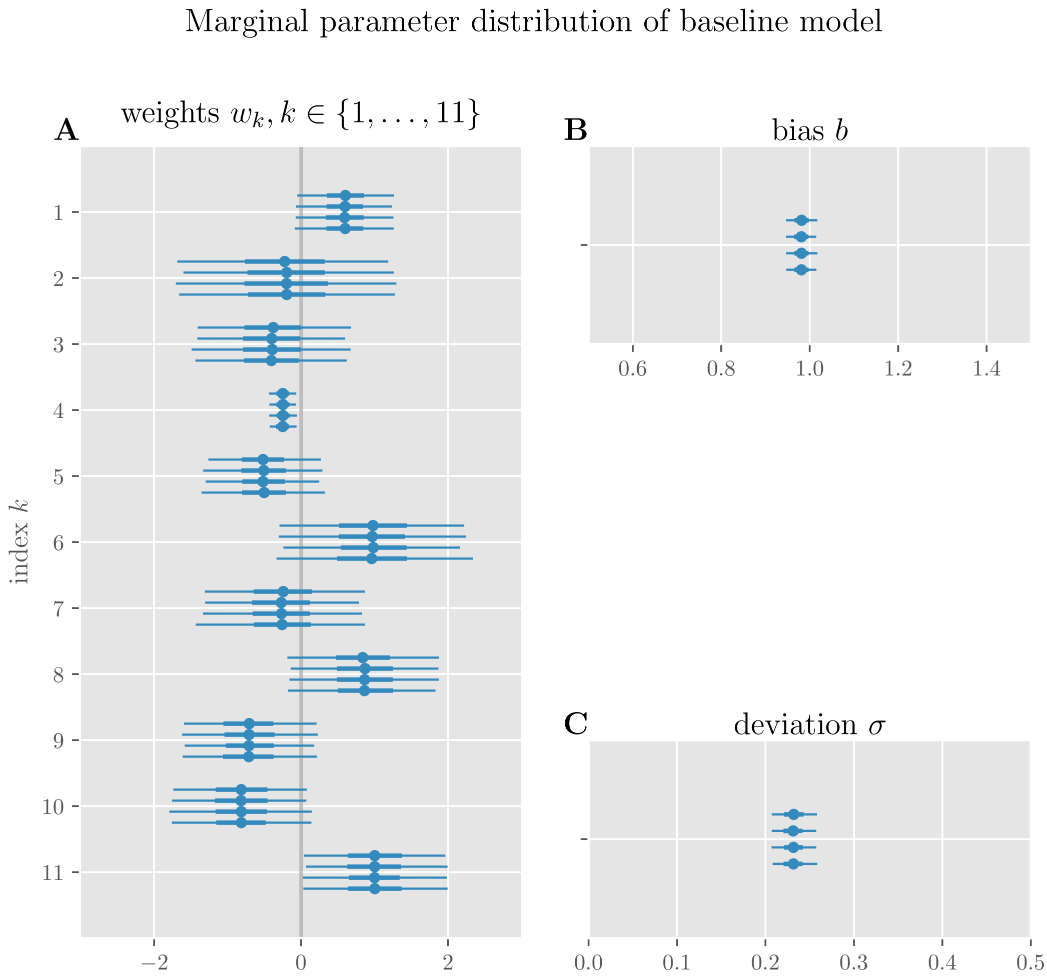
Supplementary Material: Bayesian hierarchical models can infer interpretable predictions of leaf area index from heterogeneous datasets

D.1 Figure S1D



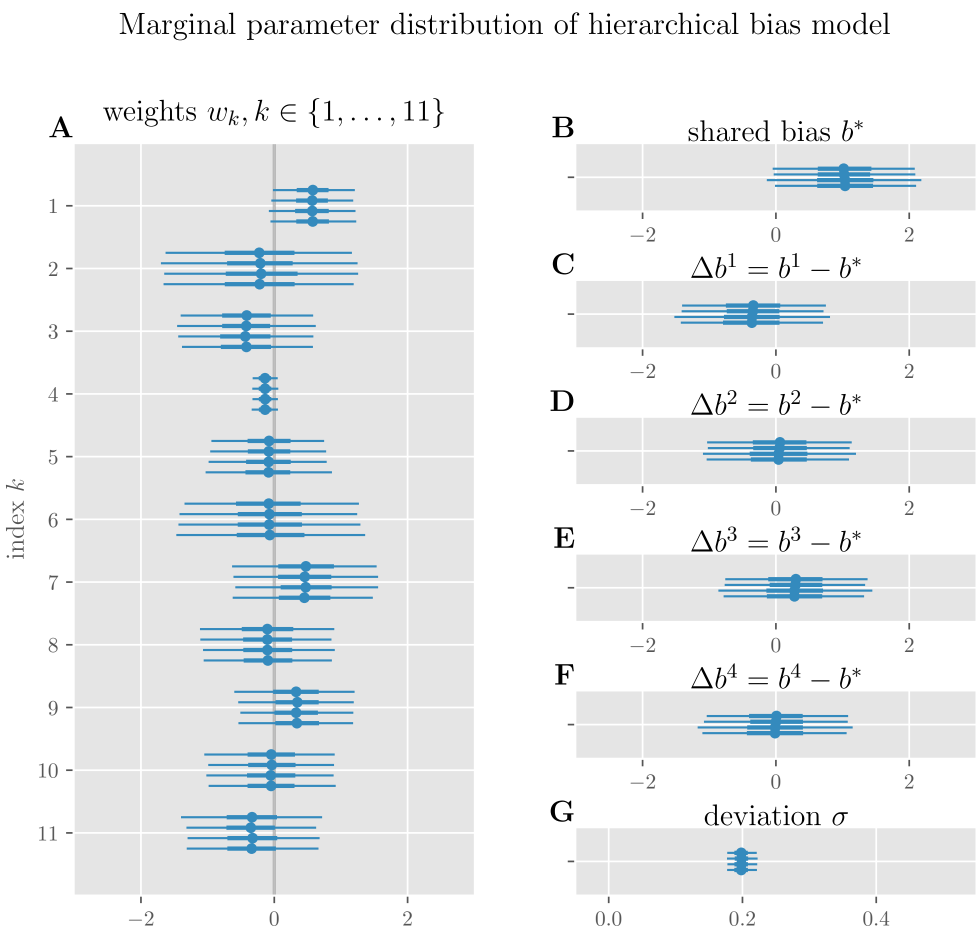
Reflectance spectra from the four datasets. Solid lines show five randomly selected spectra from each dataset. The dashed lines show the average for the respective dataset, and the gray regions show the value range (from minimum to maximum) for each wavelength.

D.2 Figure S2D



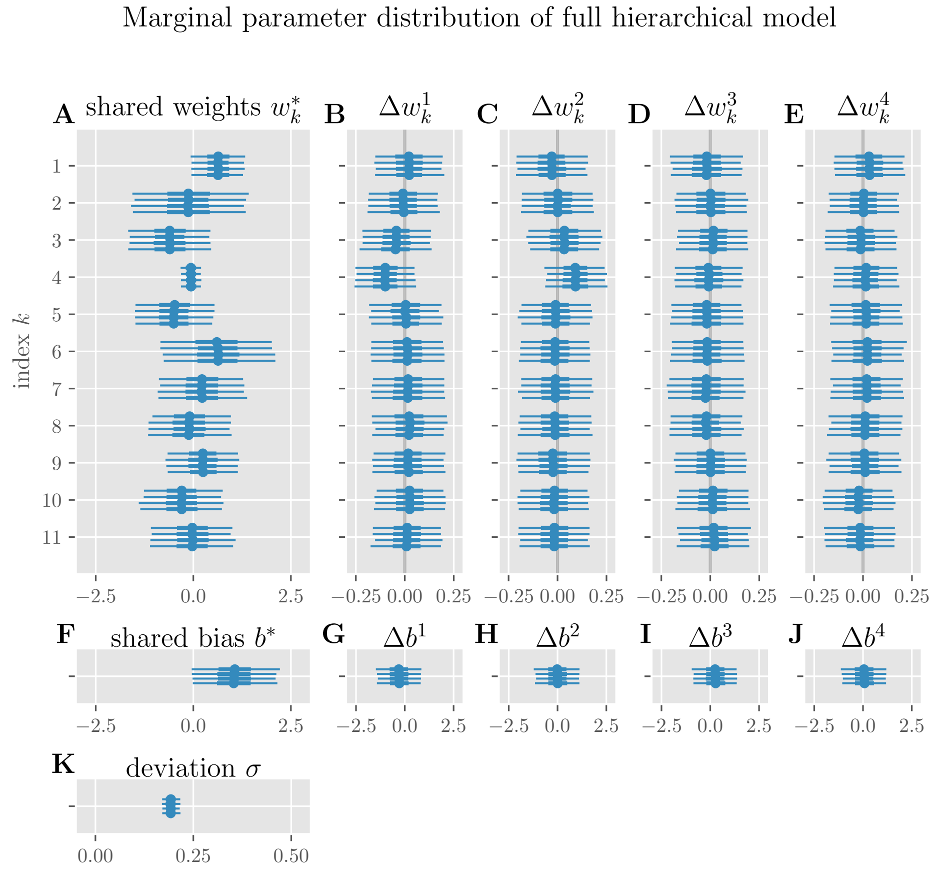
Marginal posterior distributions of all parameters of the baseline model. For each of four Markov chains, the mean (dot), the interquartile range from the 25 % to 75 % quantile (thick horizontal lines) as well as the 2.5 % to 97.5 % quantile (thin horizontal lines) are shown. For all parameters, these summary statistics of the marginal distribution are similar across all four chains, indicating convergence of the MCMC sampling scheme.

D.3 Figure S3D



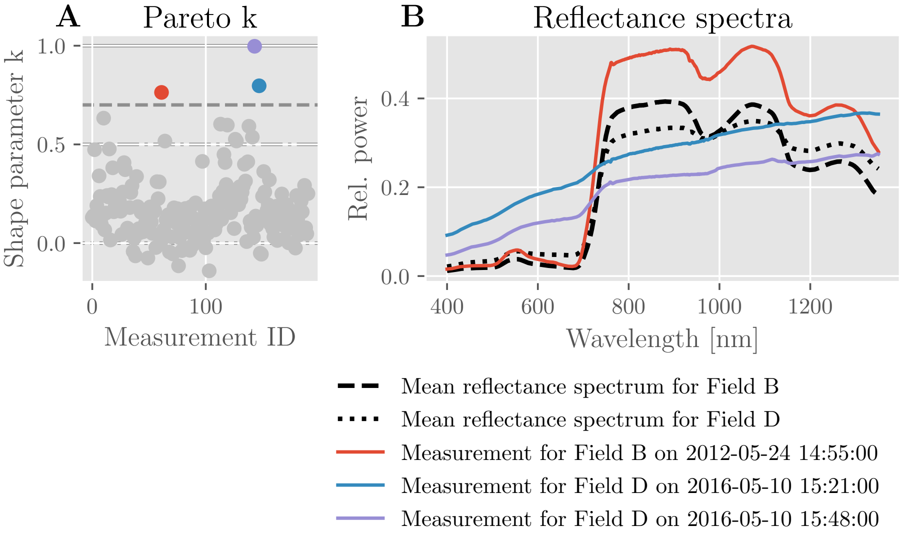
Marginal posterior distributions of all parameters of the hierarchical bias model. For each of four Markov chains, the mean (dot), the interquartile range from the 25 % to 75 % quantile (thick horizontal lines) as well as the 2.5 % to 97.5 % quantile (thin horizontal lines) are shown. (C) to (F) show the differences between the shared bias parameter b^* and the dataset-specific bias parameters b^j. For all parameters, these summary statistics of the marginal distribution are similar across all four chains, indicating convergence of the MCMC sampling scheme.

D.4 Figure S4D



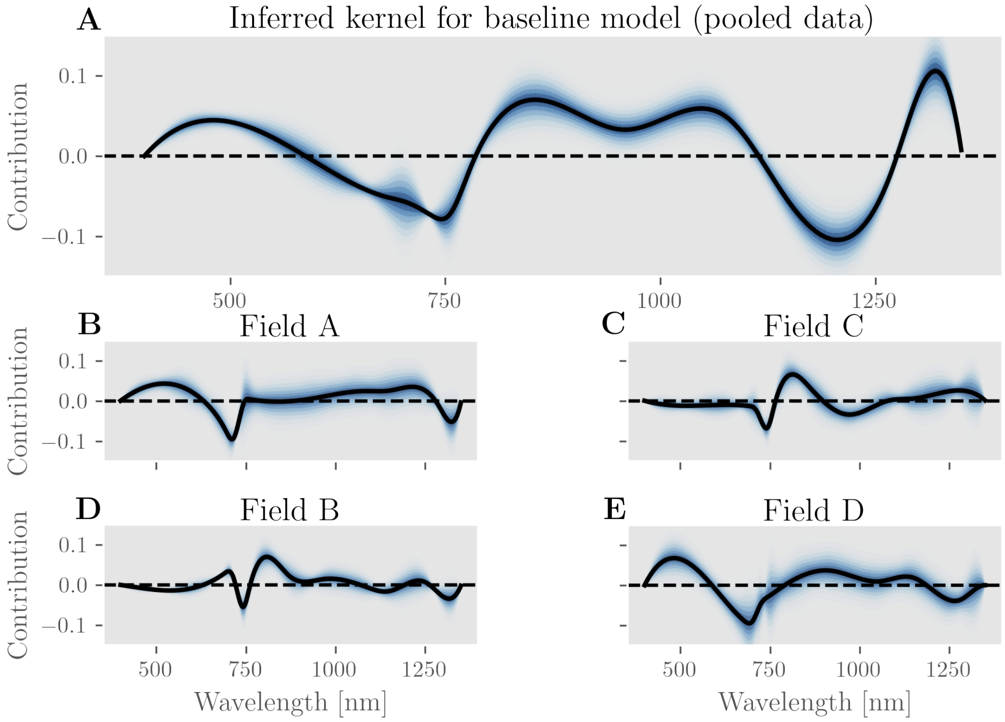
Marginal posterior distributions of all parameters of the full hierarchical model. For each of four Markov chains, the mean (dot), the interquartile range from the 25 % to 75 % quantile (thick horizontal lines) as well as the 2.5 % to 97.5 % quantile (thin horizontal lines) are shown. (B) to (E) show the differences between the shared weight parameters w_k^* and the dataset-specific weight parameters w_k^j, and (G) to (J) show the differences between the shared bias parameter b^* and the dataset-specific bias parameters b^j. For all parameters, these summary statistics of the marginal distribution are similar across all four chains, indicating convergence of the MCMC sampling scheme.

D.5 Figure S5D



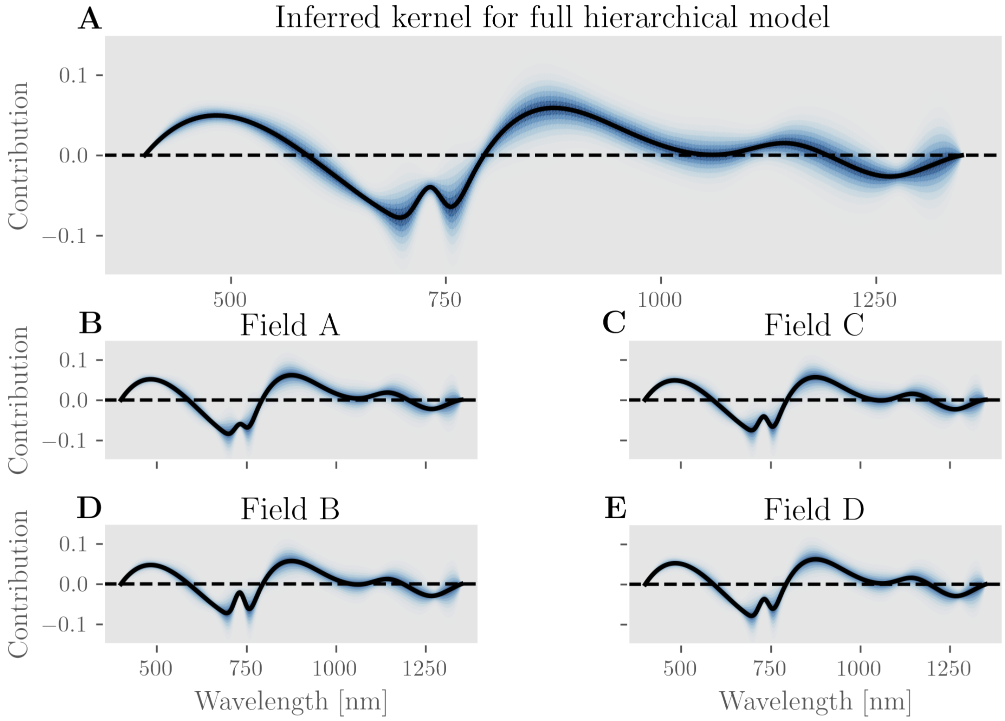
Evaluation of PSIS-LOO-CV diagnostic for full hierarchical model. (A) shows the shape parameters (Pareto k) computed for each datapoint by the PSIS-LOO-CV method for the full hierarchical model. For four measurements (color-coded), the shape parameter exceeds the critical value of 0.7 and PSIS-LOO-CV becomes unreliable. (B) shows the reflectance spectra corresponding to these four datapoints (solid lines, color-coded) as well as the mean reflectance spectra of the respective datasets (dashed lines).

D.6 Figure S6D



Inferred kernel for baseline model. (A) shows the kernel of the baseline model when fit to the entire pooled dataset. For reference, (B) through (E) show the different kernels that the baseline model would infer from each of the four datasets in isolation. We can see large, qualitative differences between the five shown kernel functions. In particular, the pooled data results in a kernel function with a sizeable dip around a wavelength of 1200 nm, which is entirely absent from any of the kernels inferred for the individual datasets. This indicates that systematic differences between the datasets might introduce spurious associations between spectral features and LAI predictions, which pose a risk for misinterpretation. This effect is avoided entirely by a full hierarchical model (see Figure S7D) and much reduced by the hierarchical bias model (see figure 6).

D.7 Figure S7D



Inferred kernel for full hierarchical model. (A) shows the shared kernel function of the full hierarchical model, and (B) through (E) show the specific kernel functions inferred for each dataset. The inferred dataset-specific kernels deviate only little from the shared kernel, yet in contrast to the baseline model in Figure S6D, there is no pronounced dip around the wavelength 1200 nm.

Bibliography

[1] World Health Organization WHO. *Naming the coronavirus disease (COVID-19) and the virus that causes it.* en. 2020. URL: `https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it` (visited on 12/06/2022).

[2] Laure Wynants et al. "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal". In: *BMJ* 369 (2020). DOI: `10.1136/bmj.m1328`. eprint: `https://www.bmj.com/content/369/bmj.m1328.full.pdf`. URL: `https://www.bmj.com/content/369/bmj.m1328`.

[3] Bhaskar Chakravorti. "Why AI Failed to Live Up to Its Potential During the Pandemic". In: *Harvard Business Review* (Mar. 2022). ISSN: 0017-8012. URL: `https://hbr.org/2022/03/why-ai-failed-to-live-up-to-its-potential-during-the-pandemic` (visited on 05/07/2022).

[4] Will Douglas Heaven. *Hundreds of AI tools have been built to catch covid. None of them helped.* en. 2021. URL: `https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/` (visited on 05/14/2022).

[5] Michael Roberts et al. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans". en. In: *Nature Machine Intelligence* 3.3 (Mar. 2021), pp. 199–217. ISSN: 2522-5839. DOI: `10.1038/s42256-021-00307-0`. URL: `https://www.nature.com/articles/s42256-021-00307-0` (visited on 05/14/2022).

[6] Christian Deker. *Kontaktlisten: Polizei fragt in mehr als 100 Fällen Daten ab.* de. 2022. URL: `https://www.zdf.de/uri/a7769257-270b-4baf-a304-951c89e9e677` (visited on 05/14/2022).

[7] Inken von Borzyskowski et al. *Data science and AI in the age of COVID-19 – report.* en. Tech. rep. London, UK: The Alan Turing Institute, 2021. URL: `https://www.turing.ac.uk/sites/default/files/2021-06/data-science-and-ai-in-the-age-of-covid_full-report_2.pdf` (visited on 05/14/2022).

[8] Shamim M. Baker, Otis W. Brawley, and Leonard S. Marks. "Effects of untreated syphilis in the negro male, 1932 to 1972: a closure comes to the Tuskegee study, 2004". eng. In: *Urology* 65.6 (June 2005), pp. 1259–1262. ISSN: 1527-9995. DOI: 10.1016/j.urology.2004.10.023.

[9] Theresa Vargas. "Guinea pigs or pioneers? How Puerto Rican women were used to test the birth control pill." en-US. In: *Washington Post* (2017). ISSN: 0190-8286. URL: https://www.washingtonpost.com/news/retropolis/wp/2017/05/09/guinea-pigs-or-pioneers-how-puerto-rican-women-were-used-to-test-the-birth-control-pill/ (visited on 05/14/2022).

[10] Whitney R Robinson, Audrey Renson, and Ashley I Naimi. "Teaching yourself about structural racism will improve your machine learning". In: *Biostatistics (Oxford, England)* 21.2 (Nov. 2019), pp. 339–344. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxz040. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7868043/ (visited on 05/14/2022).

[11] Catherine D'Ignazio and Laura Klein. *Data Feminism*. MIT Press, 2020.

[12] J. Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*. 2009.

[13] Yann LeCun, Corinna Cortes, and CJ Burges. "MNIST handwritten digit database". In: *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2 (2010).

[14] Alex Krizhevsky. *Learning multiple layers of features from tiny images.* Tech. rep. 2009.

[15] Catherine D'Ignazio. "Long read: Putting data back into context". en. In: *DataJournalism.com* (2019). URL: https://datajournalism.com/read/longreads/putting-data-back-into-context (visited on 06/03/2019).

[16] Donna Haraway. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective". In: *Feminist Studies* 14.3 (1988), pp. 575–599. ISSN: 00463663. URL: http://www.jstor.org/stable/3178066 (visited on 08/27/2022).

[17] Jill Walker Rettberg. "Situated data analysis: a new method for analysing encoded power relationships in social media platforms and apps". In: *Humanities and Social Sciences Communications* 7 (June 2020). DOI: 10.1057/s41599-020-0495-3.

[18] Heather Krause. *Data Biographies: Getting to Know Your Data*. en-US. Mar. 2017. URL: https://gijn.org/2017/03/27/data-biographies-getting-to-know-your-data/ (visited on 07/15/2019).

[19] Mary Elizabeth Luka and Mélanie Millette. "(Re)framing Big Data: Activating Situated Knowledges and a Feminist Ethics of Care in Social Media Research". In: *Social Media + Society* 4.2 (2018), p. 2056305118768297. DOI: 10.1177/2056305118768297. eprint: https://doi.org/10.1177/2056305118768297. URL: https://doi.org/10.1177/2056305118768297.

[20] Stephanie Pappas. "How COVID-19 Deaths Are Counted". en. In: *Scientific American* (2020). URL: https://www.scientificamerican.com/article/how-covid-19-deaths-are-counted1/ (visited on 02/04/2023).

[21] Catherine D'Ignazio and Lauren F. Klein. "Seven intersectional feminist principles for equitable and actionable COVID-19 data". In: *Big Data & Society* 7.2 (2020). PMID: 32802347, p. 2053951720942544. DOI: 10.1177/2053951720942544. eprint: https://doi.org/10.1177/2053951720942544. URL: https://doi.org/10.1177/2053951720942544.

[22] Mimi Onuha. *Where We Live and How We Die.* en-GB. 2016. URL: https://www.howwegettonext.com/where-we-live-and-how-we-die/ (visited on 01/07/2022).

[23] Mimi Onuoha. *An overview and exploration of the concept of missing datasets. : MimiOnuoha/missing-datasets.* original-date: 2016-02-03T16:30:28Z. 2019. URL: https://github.com/MimiOnuoha/missing-datasets (visited on 07/15/2019).

[24] Amanda Makulec. "Ten Considerations Before you Create another Chart about COVID-19". en. In: *Nightingale - The Journal of Data Visualization Society* (2020). URL: https://medium.com/nightingale/ten-considerations-before-you-create-another-chart-about-covid-19-27d3bd691be8 (visited on 05/19/2020).

[25] Lisa Charlotte Rost. "17 (or so) responsible live visualizations about the coronavirus, for you to use". en. In: *Chartable, Datawrapper GmbH* (2020). URL: https://blog.datawrapper.de/coronaviruscharts/index.html (visited on 05/19/2020).

[26] Amandalynne Paullada et al. "Data and its (dis)contents: A survey of dataset development and use in machine learning research". In: *Patterns* 2.11 (2021), p. 100336. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2021.100336. URL: https://www.sciencedirect.com/science/article/pii/S2666389921001847.

[27] Emily Denton et al. "On the genealogy of machine learning datasets: A critical history of ImageNet". In: *Big Data & Society* 8.2 (2021), p. 20539517211035955. DOI: 10.1177/20539517211035955. eprint: https://doi.org/10.1177/20539517211035955. URL: https://doi.org/10.1177/20539517211035955.

[28] David Donoho. “50 Years of Data Science”. In: *Journal of Computational and Graphical Statistics* 26.4 (Oct. 2017), pp. 745–766. ISSN: 1061-8600. DOI: 10.1080/10618600.2017.1384734. URL: https://doi.org/10.1080/10618600.2017.1384734 (visited on 07/15/2019).

[29] François Bocquet, Mario Campone, and Marc Cuggia. “The Challenges of Implementing Comprehensive Clinical Data Warehouses in Hospitals”. In: *International Journal of Environmental Research and Public Health* 19.12 (June 2022), p. 7379. ISSN: 1661-7827. DOI: 10.3390/ijerph19127379. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9223495/ (visited on 02/04/2023).

[30] Antonio Torralba and Alexei A. Efros. “Unbiased look at dataset bias”. In: *CVPR 2011*. ISSN: 1063-6919. June 2011, pp. 1521–1528. DOI: 10.1109/CVPR.2011.5995347.

[31] Matthias Ihle et al. “EPILEPSIAE – A European epilepsy database”. In: *Computer Methods and Programs in Biomedicine* 106.3 (2012), pp. 127–138. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2010.08.011.

[32] Kais Gadhoumi et al. “Seizure prediction for therapeutic devices: A review”. In: *Journal of Neuroscience Methods* 260.Supplement C (2016), pp. 270–282. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2015.06.010.

[33] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. en. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x. URL: https://www.nature.com/articles/s42256-019-0048-x (visited on 12/27/2021).

[34] Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretable Machine Learning”. In: *arXiv:1702.08608 [cs, stat]* (Feb. 2017). arXiv: 1702.08608. URL: http://arxiv.org/abs/1702.08608 (visited on 07/15/2019).

[35] Christoph Molnar. *Interpretable Machine Learning*. 2019. URL: https://christophm.github.io/interpretable-ml-book/ (visited on 06/03/2019).

[36] Patrick Schramowski et al. “Making deep neural networks right for the right scientific reasons by interacting with their explanations”. en. In: *Nature Machine Intelligence* 2.8 (2020), pp. 476–486. ISSN: 2522-5839. DOI: 10.1038/s42256-020-0212-3. URL: https://www.nature.com/articles/s42256-020-0212-3 (visited on 12/29/2021).

[37] Andrew F. Voter et al. “Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Intracranial Hemorrhage”. In: *Journal of the American College of Radiology* 18.8 (2021), pp. 1143–1152. ISSN: 1546-1440. DOI:

https://doi.org/10.1016/j.jacr.2021.03.005. URL: https://www.sciencedirect.com/science/article/pii/S1546144021002271.

[38] Marcus A. Badgeley et al. "Deep learning predicts hip fracture using confounding patient and healthcare variables". en. In: *npj Digital Medicine* 2.1 (2019), pp. 1–10. ISSN: 2398-6352. DOI: 10.1038/s41746-019-0105-1. URL: https://www.nature.com/articles/s41746-019-0105-1 (visited on 12/29/2021).

[39] Ryuji Hamamoto et al. "Application of Artificial Intelligence Technology in Oncology: Towards the Establishment of Precision Medicine". In: *Cancers* 12.12 (2020), p. 3532. ISSN: 2072-6694. DOI: 10.3390/cancers12123532. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7760590/ (visited on 12/29/2021).

[40] BreezoMeter. *The World's Most Accurate Air Quality Data — BreezoMeter.* en. 2021. URL: https://www.breezometer.com/ (visited on 12/29/2021).

[41] Charles W. Schmidt. "Into the Black Box: What Can Machine Learning Offer Environmental Health Research?" In: *Environmental Health Perspectives* 128.2 (2020), p. 022001. ISSN: 0091-6765. DOI: 10.1289/EHP5878. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7064317/ (visited on 12/27/2021).

[42] Matthew Cole. "A-level results: confusion is the result of months of inertia and years of policy". en. In: *The Conversation* (2020). URL: http://theconversation.com/a-level-results-confusion-is-the-result-of-months-of-inertia-and-years-of-policy-144260 (visited on 11/15/2020).

[43] Condé Nast. "The UK's A-level fiasco has left thousands without grades". en-GB. In: *Wired UK* (2020). ISSN: 1357-0978. URL: https://www.wired.co.uk/article/alevel-homeschool-grades-university (visited on 12/27/2021).

[44] *SCHUFA.* URL: https://www.schufa.de/en/about-us/company/schufa/schufa.jsp (visited on 01/08/2022).

[45] AlgorithmWatch. *SCHUFA, a black box: OpenSCHUFA results published.* en. 2018. URL: https://algorithmwatch.org/en/schufa-a-black-box-openschufa-results-published/ (visited on 01/08/2022).

[46] Christina Elmer et al. "Schufa: So funktioniert Deutschlands einflussreichste Auskunftei". de. In: *Der Spiegel* (Nov. 2018). ISSN: 2195-1349. URL: https://www.spiegel.de/wirtschaft/service/schufa-so-funktioniert-deutschlands-einflussreichste-auskunftei-a-1239214.html (visited on 01/08/2022).

[47] Open Knowledge Foundation Germany. *OKF DE.* de. 2022. URL: https://okfn.de (visited on 11/11/2022).

[48] Algorithm Watch and Open Knowledge Foundation Deutschland. *OpenSCHUFA*. de-DE. Tech. rep. 2020. URL: https://openschufa.de/english/ (visited on 11/14/2020).

[49] Statistisches Bundesamt DESTATIS. *Bis 2035 wird die Zahl der Menschen ab 67 Jahre um 22 % steigen*. de. 2021. URL: https://www.destatis.de/DE/Presse/Pressemitteilungen/2021/09/PD21_459_12411.html (visited on 06/11/2022).

[50] SCHUFA Holding AG. *So funktioniert der SCHUFA Score-Simulator*. DE. 2022. URL: https://www.schufa.de/themenportal/so-funktioniert-schufa-score-simulator/ (visited on 11/11/2022).

[51] Madhumita Murgia. "Emotion recognition: can AI detect human feelings from a face?" en. In: *Financial Times* (May 2021).

[52] Kate Crawford. "Time to regulate AI that interprets human emotions". en. In: *Nature* 592.7853 (Apr. 2021), pp. 167–167. DOI: 10.1038/d41586-021-00868-5. URL: https://www.nature.com/articles/d41586-021-00868-5 (visited on 11/16/2022).

[53] Ivan Manokha. *Facial analysis AI is being used in job interviews – it will probably reinforce inequality*. en. 2019. URL: http://theconversation.com/facial-analysis-ai-is-being-used-in-job-interviews-it-will-probably-reinforce-inequality-124790 (visited on 11/16/2022).

[54] Niamh Kinchin. *AI facial analysis is scientifically questionable. Should we be using it for border control?* en. 2021. URL: http://theconversation.com/ai-facial-analysis-is-scientifically-questionable-should-we-be-using-it-for-border-control-155474 (visited on 01/08/2022).

[55] Lisa Feldman Barrett et al. "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements". In: *Psychological Science in the Public Interest* 20.1 (2019). PMID: 31313636, pp. 1–68. DOI: 10.1177/1529100619832930. eprint: https://doi.org/10.1177/1529100619832930. URL: https://doi.org/10.1177/1529100619832930.

[56] Shan Jia et al. "Detection of Genuine and Posed Facial Expressions of Emotion: Databases and Methods". In: *Frontiers in Psychology* 11 (2021). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.580287. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.580287.

[57] Katharina Weitz et al. In: *tm - Technisches Messen* 86.7-8 (2019), pp. 404–412. DOI: doi:10.1515/teme-2019-0024. URL: https://doi.org/10.1515/teme-2019-0024.

[58] Juliet Cohen. "Questions of Credibility: Omissions, Discrepancies and Errors of Recall in the Testimony of Asylum Seekers". In: *International Journal of Refugee Law* 13 (July 2001). DOI: 10.1093/ijrl/13.3.293.

[59] Alex Hern. “Information commissioner warns firms over ‘emotional analysis’ technologies”. en-GB. In: *The Guardian* (Oct. 2022). ISSN: 0261-3077. URL: https://www.theguardian.com/technology/2022/oct/25/information-commissioner-warns-firms-over-emotional-analysis-technologies (visited on 10/31/2022).

[60] Zev Rosenwaks. “Artificial intelligence in reproductive medicine: a fleeting concept or the wave of the future?” eng. In: *Fertility and Sterility* 114.5 (Nov. 2020), pp. 905–907. ISSN: 1556-5653. DOI: 10.1016/j.fertnstert.2020.10.002.

[61] Michael Anis Mihdi Afnan et al. “Ethical Implementation of Artificial Intelligence to Select Embryos in In Vitro Fertilization”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* New York, NY, USA: Association for Computing Machinery, 2021, pp. 316–326. ISBN: 9781450384735. URL: https://doi.org/10.1145/3461702.3462589.

[62] J. Savulescu, I. Chalmers, and J. Blunt. “Are research ethics committees behaving unethically? Some suggestions for improving performance and accountability.” In: *BMJ : British Medical Journal* 313.7069 (Nov. 1996), pp. 1390–1393. ISSN: 0959-8138. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2352884/ (visited on 01/08/2022).

[63] Derek Parfit. *Reasons and Persons.* Oxford, UK: Oxford University Press, 1984.

[64] Julian Savulescu and Guy Kahane. “Understanding Procreative Beneficence”. In: *Oxford Handbooks Online OP* (2016). DOI: 10.1093/oxfordhb/9780199981878.013.26.

[65] Ute Schmid and Bettina Finzel. “Mutual Explanations for Cooperative Decision Making in Medicine”. en. In: *KI - Künstliche Intelligenz* 34.2 (June 2020), pp. 227–233. ISSN: 1610-1987. DOI: 10.1007/s13218-020-00633-2. URL: https://doi.org/10.1007/s13218-020-00633-2 (visited on 03/03/2023).

[66] Sebastian Bruckert, Bettina Finzel, and Ute Schmid. “The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions”. In: *Frontiers in Artificial Intelligence* 3 (2020). ISSN: 2624-8212. DOI: 10.3389/frai.2020.507973. URL: https://www.frontiersin.org/articles/10.3389/frai.2020.507973.

[67] Emanuel Slany et al. “CAIPI in Practice: Towards Explainable Interactive Medical Image Classification”. en. In: *Artificial Intelligence Applications and Innovations. AIAI 2022 IFIP WG 12.5 International Workshops.* Ed. by Ilias Maglogiannis et al. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, 2022, pp. 389–400. ISBN: 9783031083419. DOI: 10.1007/978-3-031-08341-9_31.

[68] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. “Machine Learning Interpretability: A Survey on Methods and Metrics”. In: *Electronics* 8.8 (2019). ISSN: 2079-9292. DOI: 10.3390/electronics8080832. URL: https://www.mdpi.com/2079-9292/8/8/832.

[69] Cynthia Rudin et al. *Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges*. 2021. arXiv: 2103.11251 [cs.LG].

[70] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[71] W. James Murdoch et al. “Definitions, methods, and applications in interpretable machine learning”. In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080. ISSN: 0027-8424. DOI: 10.1073/pnas.1900654116. eprint: https://www.pnas.org/content/116/44/22071.full.pdf. URL: https://www.pnas.org/content/116/44/22071.

[72] Riccardo Guidotti et al. “A Survey of Methods for Explaining Black Box Models”. In: *ACM Comput. Surv.* 51.5 (Aug. 2018). ISSN: 0360-0300. DOI: 10.1145/3236009. URL: https://doi.org/10.1145/3236009.

[73] Christoph Molnar et al. “Pitfalls to Avoid when Interpreting Machine Learning Models”. In: *ArXiv* abs/2007.04131 (2020).

[74] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. “iml: An R package for Interpretable Machine Learning”. In: *Journal of Open Source Software* 3.26 (2018), p. 786. DOI: 10.21105/joss.00786. URL: https://doi.org/10.21105/joss.00786.

[75] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org.

[76] James Wexler et al. “The what-if tool: Interactive probing of machine learning models”. In: *IEEE transactions on visualization and computer graphics* 26.1 (2019), pp. 56–65.

[77] Janis Klaise et al. “Alibi Explain: Algorithms for Explaining Machine Learning Models”. In: *Journal of Machine Learning Research* 22.181 (2021), pp. 1–7. ISSN: 1533-7928. URL: http://jmlr.org/papers/v22/21-0017.html (visited on 01/08/2022).

[78] Patrick Hall et al. “Machine learning interpretability with h2o driverless ai”. In: *H2O. ai* (2017).

[79] Przemyslaw Biecek. “DALEX: Explainers for Complex Predictive Models in R”. In: *Journal of Machine Learning Research* 19.84 (2018), pp. 1–5. ISSN: 1533-7928. URL: http://jmlr.org/papers/v19/18-416.html (visited on 01/08/2022).

[80] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. *Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges.* 2020. arXiv: `2010.09337 [stat.ML]`.

[81] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. "Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (May 2020). arXiv: 2004.11440, pp. 1–26. ISSN: 2573-0142, 2573-0142. DOI: `10.1145/3392878`. URL: `http://arxiv.org/abs/2004.11440` (visited on 08/09/2020).

[82] Patrick Hall and Navdeep Gill. *An Introduction to Machine Learning Interpretability.* en. 1st. O'Reilly Media, Inc., 2018. URL: `https://pages.dataiku.com/hubfs/ML-interperatability.pdf`.

[83] Gary Marcus and Ernest Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust.* en-US. 1st. USA: Pantheon Books, 2019. ISBN: 9781524748258. URL: `https://www.penguinrandomhouse.com/books/603982/rebooting-ai-by-gary-marcus-and-ernest-davis/` (visited on 03/08/2020).

[84] Tim Miller. "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *arXiv:1706.07269 [cs]* (June 2017). arXiv: 1706.07269. URL: `http://arxiv.org/abs/1706.07269` (visited on 07/15/2019).

[85] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. "Examples are not enough, learn to criticize! Criticism for Interpretability". en. In: (2016), p. 9.

[86] Zachary C. Lipton. "The Mythos of Model Interpretability". In: *arXiv:1606.03490 [cs, stat]* (June 2016). arXiv: 1606.03490. URL: `http://arxiv.org/abs/1606.03490` (visited on 07/15/2019).

[87] Dina Mardaoui and Damien Garreau. "An Analysis of LIME for Text Data". en. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics.* PMLR, Mar. 2021, pp. 3493–3501. URL: `https://proceedings.mlr.press/v130/mardaoui21a.html` (visited on 03/03/2023).

[88] Ludwig Schallner et al. "Effect of Superpixel Aggregation on Explanations in LIME – A Case Study with Biological Data". en. In: *Machine Learning and Knowledge Discovery in Databases.* Ed. by Peggy Cellier and Kurt Driessens. Communications in Computer and Information Science. Cham: Springer International Publishing, 2020, pp. 147–158. ISBN: 9783030438234. DOI: `10.1007/978-3-030-43823-4_13`.

[89] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously". In: *arXiv:1801.01489 [stat]* (2019). arXiv: 1801.01489, version: 5. URL: `http://arxiv.org/abs/1801.01489` (visited on 12/10/2020).

[90] Jeff Larson Julia Angwin. *Machine Bias.* en. text/html. May 2016. URL: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing` (visited on 07/15/2019).

[91] Julia Angwin Jeff Larson. *How We Analyzed the COMPAS Recidivism Algorithm*. en. text/html. May 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (visited on 07/15/2019).

[92] *propublica/compas-analysis*. original-date: 2016-05-21T03:36:35Z. Nov. 2019. URL: https://github.com/propublica/compas-analysis.

[93] Tim Brennan, William Dieterich, and Beate Ehret. "Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System". In: *Criminal Justice and Behavior* 36.1 (2009), pp. 21–40. DOI: 10.1177/0093854808326545. eprint: https://doi.org/10.1177/0093854808326545. URL: https://doi.org/10.1177/0093854808326545.

[94] Michael Tonry. "Legal and Ethical Issues in the Prediction of Recidivism". In: *Federal Sentencing Reporter* 26 (2014), p. 167. URL: https://scholarship.law.umn.edu/faculty_articles/525.

[95] Cynthia Rudin, Caroline Wang, and Beau Coker. "The Age of Secrecy and Unfairness in Recidivism Prediction". In: *Harvard Data Science Review* 2.1 (Mar. 31, 2020). https://hdsr.mitpress.mit.edu/pub/7z10o269. DOI: 10.1162/99608f92.6ed64b30. URL: https://hdsr.mitpress.mit.edu/pub/7z10o269.

[96] Lieven Billiet, Sabine Van Huffel, and Vanya Van Belle. "Interval Coded Scoring extensions for larger problems". In: *2017 IEEE Symposium on Computers and Communications (ISCC)*. 2017, pp. 198–203. DOI: 10.1109/ISCC.2017.8024529.

[97] Lieven Billiet, Sabine Van Huffel, and Vanya Van Belle. "Interval Coded Scoring: a toolbox for interpretable scoring systems". en. In: *PeerJ Computer Science* 4 (Apr. 2018), e150. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.150. URL: https://peerj.com/articles/cs-150 (visited on 11/12/2022).

[98] Ángel Alexander Cabrera et al. "FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning". In: *arXiv:1904.05419 [cs, stat]* (Sept. 2019). arXiv: 1904.05419. URL: http://arxiv.org/abs/1904.05419 (visited on 04/11/2020).

[99] Leo Breiman. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". en. In: *Statistical Science* 16.3 (Aug. 2001), pp. 199–231. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/ss/1009213726. URL: https://projecteuclid.org/euclid.ss/1009213726 (visited on 07/15/2019).

[100] Lesia Semenova, Cynthia Rudin, and Ronald Parr. *A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning*. 2021. arXiv: 1908.01755 [cs.LG].

[101] Daniel D. Lee and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization". En. In: *Nature* 401.6755 (1999), p. 788. ISSN: 1476-4687. DOI: 10.1038/44565. (Visited on 11/27/2017).

[102] Pentti Paatero and Unto Tapper. “Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values”. en. In: *Environmetrics* 5.2 (June 1994), pp. 111–126. ISSN: 1099-095X. DOI: 10.1002/env.3170050203. URL: http://onlinelibrary.wiley.com/doi/10.1002/env.3170050203/abstract (visited on 03/15/2018).

[103] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 0035-9246. URL: https://www.jstor.org/stable/2346178 (visited on 03/21/2019).

[104] Mei-Sing Ong et al. “Population-Level Evidence for an Autoimmune Etiology of Epilepsy”. In: *JAMA neurology* 71.5 (2014), pp. 569–574. ISSN: 2168-6149. DOI: 10.1001/jamaneurol.2014.188. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4324719/ (visited on 11/22/2022).

[105] C. P. Ed. Panayiotopoulos. *Atlas of Epilepsies*. Springer, 2010. (Visited on 12/04/2017).

[106] Sriram Ramgopal et al. “Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy”. en. In: *Epilepsy & Behavior* 37 (2014), pp. 291–307. ISSN: 1525-5050. DOI: 10.1016/j.yebeh.2014.06.023. URL: http://www.sciencedirect.com/science/article/pii/S1525505014002297 (visited on 03/29/2020).

[107] Florian Mormann et al. “Seizure prediction: the long and winding road”. eng. In: *Brain: A Journal of Neurology* 130.Pt 2 (2007), pp. 314–333. ISSN: 1460-2156. DOI: 10.1093/brain/awl241.

[108] Turkey N. Alotaiby et al. “EEG seizure detection and prediction algorithms: a survey”. en. In: *EURASIP Journal on Advances in Signal Processing* 2014.1 (Dec. 2014), p. 183. ISSN: 1687-6180. DOI: 10.1186/1687-6180-2014-183. URL: https://link.springer.com/article/10.1186/1687-6180-2014-183 (visited on 03/09/2018).

[109] Levin Kuhlmann et al. “Seizure prediction: ready for a new era”. English. In: *Nature Reviews Neurology* 14 (2018), pp. 618–630. ISSN: 1759-4766. DOI: 10.1038/s41582-018-0055-2.

[110] Mark J Cook et al. “Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study”. In: 12.6 (2013), pp. 563–571. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(13)70075-9. (Visited on 12/04/2017).

[111] Mayela Zamora et al. “DyNeuMo Mk-1: Design and pilot validation of an investigational motion-adaptive neurostimulator with integrated chronotherapy”. In: *Experimental neurology* 351 (2022), p. 113977. ISSN: 0014-4886. DOI: 10.1016/j.expneurol.2022.113977. URL: https://europepmc.org/articles/PMC7612891.

[112] Florian Mormann and Ralph G. Andrzejak. "Seizure prediction: making mileage on the long and winding road". In: *Brain* 139.6 (2016), pp. 1625–1627. ISSN: 0006-8950. DOI: 10.1093/brain/aww091. (Visited on 12/04/2017).

[113] Aina Puce and Matti S. Hämäläinen. "A Review of Issues Related to Data Acquisition and Analysis in EEG/MEG Studies". In: *Brain Sciences* 7.6 (May 2017), p. 58. ISSN: 2076-3425. DOI: 10.3390/brainsci7060058. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5483631/ (visited on 03/03/2023).

[114] Levin Kuhlmann et al. "Epilepsyecosystem.org: crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG". In: *Brain* 141.9 (Aug. 2018), pp. 2619–2630. ISSN: 0006-8950. DOI: 10.1093/brain/awy210. eprint: http://oup.prod.sis.lan/brain/article-pdf/141/9/2619/25590596/awy210.pdf. URL: https://doi.org/10.1093/brain/awy210.

[115] Mojtaba Bandarabadi et al. "Epileptic seizure prediction using relative spectral power features". eng. In: *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 126.2 (2015), pp. 237–248. ISSN: 1872-8952. DOI: 10.1016/j.clinph.2014.05.022.

[116] Khansa Rasheed et al. "Machine Learning for Predicting Epileptic Seizures Using EEG Signals: A Review". In: *arXiv:2002.01925 [cs, eess, q-bio]* (Feb. 2020). arXiv: 2002.01925 version: 1. URL: http://arxiv.org/abs/2002.01925 (visited on 04/26/2020).

[117] Nhan Duy Truong et al. "A Generalised Seizure Prediction with Convolutional Neural Networks for Intracranial and Scalp Electroencephalogram Data Analysis". In: *arXiv:1707.01976 [cs]* (Dec. 2017). arXiv: 1707.01976. URL: http://arxiv.org/abs/1707.01976 (visited on 04/26/2020).

[118] Ramy Hussein et al. "Human Intracranial EEG Quantitative Analysis and Automatic Feature Learning for Epileptic Seizure Prediction". In: *arXiv:1904.03603 [cs, q-bio]* (Apr. 2019). arXiv: 1904.03603. URL: http://arxiv.org/abs/1904.03603 (visited on 04/26/2020).

[119] Kostas M. Tsiouris et al. "A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals". en. In: *Computers in Biology and Medicine* 99 (Aug. 2018), pp. 24–37. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2018.05.019. URL: http://www.sciencedirect.com/science/article/pii/S001048251830132X (visited on 04/26/2020).

[120] Imene Jemal et al. "An Interpretable Deep Learning Classifier for Epileptic Seizure Prediction Using EEG Data". In: *IEEE Access* 10 (2022), pp. 60141–60150. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3176367.

[121] Valentin Gabeff et al. “Interpreting deep learning models for epileptic seizure detection on EEG signals”. In: *Artificial Intelligence in Medicine* 117 (2021), p. 102084. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2021.102084. URL: https://www.sciencedirect.com/science/article/pii/S0933365721000774.

[122] Michele Lo Giudice et al. “Permutation Entropy-Based Interpretability of Convolutional Neural Network Models for Interictal EEG Discrimination of Subjects with Epileptic Seizures vs. Psychogenic Non-Epileptic Seizures”. en. In: *Entropy* 24.1 (2022), p. 102. ISSN: 1099-4300. DOI: 10.3390/e24010102. URL: https://www.mdpi.com/1099-4300/24/1/102.

[123] Mauro F. Pinto et al. “A personalized and evolutionary algorithm for interpretable EEG epilepsy seizure prediction”. en. In: *Scientific Reports* 11.1 (2021), p. 3415. ISSN: 2045-2322. DOI: 10.1038/s41598-021-82828-7. URL: https://www.nature.com/articles/s41598-021-82828-7 (visited on 01/22/2022).

[124] Oleg E. Karpov et al. “Detecting epileptic seizures using machine learning and interpretable features of human EEG”. en. In: *The European Physical Journal Special Topics* (2022). ISSN: 1951-6401. DOI: 10.1140/epjs/s11734-022-00714-3. URL: https://doi.org/10.1140/epjs/s11734-022-00714-3 (visited on 11/23/2022).

[125] Pia Anttila et al. “Source identification of bulk wet deposition in Finland by positive matrix factorization”. In: *Atmospheric Environment* 29.14 (Jan. 1995), pp. 1705–1718. ISSN: 1352-2310. DOI: 10.1016/1352-2310(94)00367-T. URL: http://www.sciencedirect.com/science/article/pii/135223109400367T (visited on 03/16/2018).

[126] Yang Zheng et al. “Epileptic seizure prediction using phase synchronization based on bivariate empirical mode decomposition”. In: *Clinical Neurophysiology* 125.6 (2014), pp. 1104–1111. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2013.09.047. (Visited on 12/04/2017).

[127] Levin Kuhlmann et al. “Patient-specific bivariate-synchrony-based seizure prediction for short prediction horizons”. English. In: *Epilepsy Research* 91.2-3 (Oct. 2010), pp. 214–231. ISSN: 0920-1211. DOI: 10.1016/j.eplepsyres.2010.07.014.

[128] Nhan Truong et al. “Epileptic Seizure Forecasting with Generative Adversarial Networks”. In: *IEEE Access* PP (Sept. 2019), pp. 1–1. DOI: 10.1109/ACCESS.2019.2944691.

[129] Ardalan Aarabi and Bin He. “Seizure prediction in hippocampal and neocortical epilepsy using a model-based approach”. In: *Clinical Neurophysiology* 125.5 (2014), pp. 930–940. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2013.10.051. (Visited on 12/04/2017).

[130] Philippa J. Karoly et al. “The circadian profile of epilepsy improves seizure forecasting”. English. In: *Brain* 140.8 (Aug. 2017), pp. 2169–2182. ISSN: 0006-8950. DOI: 10.1093/brain/awx173.

[131] Jean Gotman. “A few thoughts on “What is a seizure?””. In: *Epilepsy & behavior : E&B* 22.Suppl 1 (Dec. 2011), S2–S3. ISSN: 1525-5050. DOI: 10.1016/j.yebeh.2011.08.025. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3753284/ (visited on 03/19/2018).

[132] Yun Park et al. “Seizure prediction with spectral power of EEG using cost-sensitive support vector machines”. eng. In: *Epilepsia* 52.10 (2011), pp. 1761–1770. ISSN: 1528-1167. DOI: 10.1111/j.1528-1167.2011.03138.x.

[133] Benjamin H. Brinkmann et al. “Forecasting Seizures Using Intracranial EEG Measures and SVM in Naturally Occurring Canine Epilepsy”. In: *PLOS ONE* 10.8 (2015), e0133900. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0133900. (Visited on 12/04/2017).

[134] Leonidas D. Iasemidis et al. “Phase space topography and the Lyapunov exponent of electrocorticograms in partial seizures”. en. In: *Brain Topography* 2.3 (1990), pp. 187–201. ISSN: 0896-0267, 1573-6792. DOI: 10.1007/BF01140588. (Visited on 12/04/2017).

[135] M Paulus et al. “Synchronization and information flow in EEGs of epileptic patients - IEEE Journals & Magazine”. In: 20.5 (2001), pp. 65–71. DOI: 10.1109/51.956821. (Visited on 12/04/2017).

[136] R. Steuer et al. “Entropy and complexity analysis of intracranially recorded eeg”. In: *International Journal of Bifurcation and Chaos* 14.02 (2004), pp. 815–823. ISSN: 0218-1274. DOI: 10.1142/S021812740400948X. (Visited on 12/04/2017).

[137] Nitesh V. Chawla et al. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357.

[138] Andreas Schulze-Bonhage et al. “Views of patients with epilepsy on seizure prediction devices”. eng. In: *Epilepsy & Behavior: E&B* 18.4 (Aug. 2010), pp. 388–396. ISSN: 1525-5069. DOI: 10.1016/j.yebeh.2010.05.008.

[139] Mike X Cohen. *Analyzing Neural Time Series Data: Theory and Practice.* The MIT Press, 2014. (Visited on 12/04/2017).

[140] Nicolas Gillis. “The Why and How of Nonnegative Matrix Factorization — Regularization, Optimization, Kernels, and Support Vector Machines — Taylor & Francis Group”. In: *Regularization, Optimization, Kernels, and Support Vector Machines.* 1st Edition. New York: Chapman and Hall/CRC, 2014. ISBN: 978-1-4822-4140-2. URL: https://www.taylorfrancis.com/books/e/9781482241402/chapters/10.1201%2Fb17558-12 (visited on 03/08/2018).

[141] M. Rajapakse and L. Wyse. “NMF vs ICA for face recognition”. In: *3rd International Symposium on Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the.* Vol. 2. Sept. 2003, 605–610 Vol.2. DOI: 10.1109/ISPA.2003.1296348.

[142] David Guillamet and Jordi Vitrià. “Non-negative Matrix Factorization for Face Recognition”. en. In: *Topics in Artificial Intelligence.* Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2002, pp. 336–344. ISBN: 978-3-540-00011-2 978-3-540-36079-7. DOI: 10.1007/3-540-36079-4_29. URL: https://link.springer.com/chapter/10.1007/3-540-36079-4_29 (visited on 03/08/2018).

[143] S. Traitruengsakul et al. “Automatic localization of epileptic spikes in eegs of children with infantile spasms”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* Aug. 2015, pp. 6194–6197. DOI: 10.1109/EMBC.2015.7319807.

[144] Maxime O. Baud et al. “Unsupervised Learning of Spatiotemporal Interictal Discharges in Focal Epilepsy”. eng. In: *Neurosurgery* (Oct. 2017). ISSN: 1524-4040. DOI: 10.1093/neuros/nyx480.

[145] Christopher Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006. (Visited on 12/04/2017).

[146] Karlijn J. van Stralen et al. “Diagnostic methods I: sensitivity, specificity, and other measures of accuracy”. In: *Kidney International* 75.12 (2009), pp. 1257–1263. ISSN: 0085-2538. DOI: https://doi.org/10.1038/ki.2009.92.

[147] Hoameng Ung et al. “Intracranial EEG fluctuates over months after implanting electrodes in human brain.” In: *Department of Neurosurgery Faculty Papers* 14.5 (Sept. 2017), p. 056011. URL: https://jdc.jefferson.edu/neurosurgeryfp/100.

[148] Brian Litt et al. “Epileptic Seizures May Begin Hours in Advance of Clinical Onset: A Report of Five Patients”. In: *Neuron* 30.1 (2001), pp. 51–64. ISSN: 0896-6273. DOI: https://doi.org/10.1016/S0896-6273(01)00262-8. URL: http://www.sciencedirect.com/science/article/pii/S0896627301002628.

[149] Leonhard Held, Michael Höhle, and Mathias Hofmann. “A statistical framework for the analysis of multivariate infectious disease surveillance counts”. en. In: *Statistical Modelling* 5.3 (Oct. 2005), pp. 187–199. ISSN: 1471-082X. DOI: 10.1191/1471082X05st098oa. URL: https://doi.org/10.1191/1471082X05st098oa (visited on 01/27/2023).

[150] Chelsea S. Lutz et al. “Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples”. In: *BMC Public Health* 19.1 (Dec. 2019), p. 1659. ISSN: 1471-2458. DOI: 10.1186/s12889-019-7966-8. URL: https://doi.org/10.1186/s12889-019-7966-8 (visited on 01/27/2023).

[151] Stephen S. Morse et al. “Prediction and prevention of the next pandemic zoonosis”. eng. In: *Lancet (London, England)* 380.9857 (Dec. 2012), pp. 1956–1965. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(12)61684-5.

[152] Ting-Wu Chuang et al. “Epidemiological Characteristics and Space-Time Analysis of the 2015 Dengue Outbreak in the Metropolitan Region of Tainan City, Taiwan”. In: *International Journal of Environmental Research and Public Health* 15.3 (2018). ISSN: 1660-4601. DOI: 10.3390/ijerph15030396. URL: https://www.mdpi.com/1660-4601/15/3/396.

[153] Suleman Atique et al. “Investigating spatio-temporal distribution and diffusion patterns of the dengue outbreak in Swat, Pakistan”. eng. In: *Journal of Infection and Public Health* 11.4 (2018), pp. 550–557. ISSN: 1876-035X. DOI: 10.1016/j.jiph.2017.12.003.

[154] Daniel Adyro Martínez-Bello, Antonio López-Quílez, and Alexander Torres Prieto. “Spatio-Temporal Modeling of Zika and Dengue Infections within Colombia”. In: *International Journal of Environmental Research and Public Health* 15.7 (2018). ISSN: 1660-4601. DOI: 10.3390/ijerph15071376. URL: https://www.mdpi.com/1660-4601/15/7/1376.

[155] André Victor Ribeiro Amaral, Jonatan A. González, and Paula Moraga. “Spatio-temporal modeling of infectious diseases by integrating compartment and point process models”. en. In: *Stochastic Environmental Research and Risk Assessment* (Dec. 2022). ISSN: 1436-3259. DOI: 10.1007/s00477-022-02354-4. URL: https://doi.org/10.1007/s00477-022-02354-4 (visited on 01/25/2023).

[156] Micaela Elvira Martinez. “The calendar of epidemics: Seasonal cycles of infectious diseases”. In: *PLOS Pathogens* 14.11 (Nov. 2018), pp. 1–15. DOI: 10.1371/journal.ppat.1007327. URL: https://doi.org/10.1371/journal.ppat.1007327.

[157] N. Kronfeld-Schor et al. “Drivers of Infectious Disease Seasonality: Potential Implications for COVID-19”. In: *Journal of Biological Rhythms* 36.1 (2021). PMID: 33491541, pp. 35–54. DOI: 10.1177/0748730420987322. eprint: https://doi.org/10.1177/0748730420987322. URL: https://doi.org/10.1177/0748730420987322.

[158] Felix Günther et al. “Nowcasting the COVID-19 pandemic in Bavaria”. eng. In: *Biometrical Journal. Biometrische Zeitschrift* 63.3 (Mar. 2021), pp. 490–502. ISSN: 1521-4036. DOI: 10.1002/bimj.202000112.

[159] Jan van de Kassteele, Paul H. C. Eilers, and Jacco Wallinga. “Nowcasting the Number of New Symptomatic Cases During Infectious Disease Outbreaks Using Constrained P-spline Smoothing”. en-US. In: *Epidemiology* 30.5 (Sept. 2019), p. 737. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000001050. URL: https://journals.lww.com/epidem/Fulltext/2019/09000/Nowcasting_

the_Number_of_New_Symptomatic_Cases.16.aspx (visited on 01/25/2023).

[160] Tjibbe Donker et al. "Nowcasting pandemic influenza A/H1N1 2009 hospitalizations in the Netherlands". eng. In: *European Journal of Epidemiology* 26.3 (Mar. 2011), pp. 195–201. ISSN: 1573-7284. DOI: 10.1007/s10654-011-9566-5.

[161] Michael Höhle and Matthias an der Heiden. "Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011". eng. In: *Biometrics* 70.4 (Dec. 2014), pp. 993–1002. ISSN: 1541-0420. DOI: 10.1111/biom.12194.

[162] D. Faensen et al. "SurvNet@RKI–a multistate electronic reporting system for communicable diseases." In: *Eurosurveillance: bulletin européen sur les maladies transmissibles = European communicable disease bulletin.* 11.4 (2006), pp. 100–103. ISSN: 15607917. DOI: 10.2807/esm.11.04.00614-en.

[163] European Centre for Disease Prevention and Control. *Homepage — European Centre for Disease Prevention and Control.* en. 2022. URL: https://www.ecdc.europa.eu/en (visited on 01/25/2023).

[164] Eurostat. *Principles and Characteristics - NUTS - Nomenclature of territorial units for statistics.* 2022. URL: https://ec.europa.eu/eurostat/web/nuts/principles-and-characteristics (visited on 01/25/2023).

[165] Sciensano. *Epistat – COVID-19 Monitoring.* 2022. URL: https://epistat.sciensano.be/covid/ (visited on 09/26/2022).

[166] Welzijn en Sport Ministerie van Volksgezondheid. *Overview municipalities — Coronavirus Dashboard — Government.nl.* nl-NL. 2022. URL: https://coronadashboard.government.nl/veelgestelde-vragen (visited on 09/26/2022).

[167] Statbel (Directorate-general Statistics - Statistics Belgium). *Population by place of residence, nationality (Belgian/non-Belgian), marital status, age and gender.* 2022. URL: https://bestat.statbel.fgov.be/bestat/crosstable.xhtml?datasource=65ee413b-3859-4c6f-a847-09b631766fa7 (visited on 01/25/2023).

[168] Centraal Bureau voor de Statistiek. *Population dynamics; birth, death and migration per region.* nl. 2022. URL: https://opendata.cbs.nl/statline/#/CBS/en/dataset/37259eng/table?ts=1674637316026 (visited on 01/25/2023).

[169] Krzysztof Rzasa and Mateusz Ciski. "Influence of the Demographic, Social, and Environmental Factors on the COVID-19 Pandemic-Analysis of the Local Variations Using Geographically Weighted Regression". eng. In: *International Journal of Environmental Research and Public Health* 19.19 (2022), p. 11881. ISSN: 1660-4601. DOI: 10.3390/ijerph191911881.

[170] Heyuan You, Xi Wu, and Xuxu Guo. "Distribution of COVID-19 Morbidity Rate in Association with Social and Economic Factors in Wuhan, China: Implications for Urban Development". eng. In: *International Journal of Environmental Research and Public Health* 17.10 (2020), p. 3417. ISSN: 1660-4601. DOI: `10.3390/ijerph17103417`.

[171] Nushrat Nazia, Jane Law, and Zahid Ahmad Butt. "Spatiotemporal clusters and the socioeconomic determinants of COVID-19 in Toronto neighbourhoods, Canada". eng. In: *Spatial and Spatio-Temporal Epidemiology* 43 (Nov. 2022), p. 100534. ISSN: 1877-5853. DOI: `10.1016/j.sste.2022.100534`.

[172] Gordon Pipa. *Analyzing tweets to predict flu epidemics.* en-US. Aug. 2017. URL: `https://www.ibm.com/blogs/client-voices/analyzing-tweets-predict-flu-epidemics/` (visited on 01/29/2019).

[173] Derek Ruths and Jürgen Pfeffer. "Social media for large studies of behavior". en. In: *Science* 346.6213 (Nov. 2014), pp. 1063–1064. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.346.6213.1063`. URL: `https://science.sciencemag.org/content/346/6213/1063` (visited on 07/15/2019).

[174] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think.* Big data: A revolution that will transform how we live, work, and think. Boston, MA: Houghton Mifflin Harcourt, 2013. ISBN: 9780544002692 9780544002937.

[175] The Publications Office of the European Union. *What is open data — data.europa.eu.* 2023. URL: `https://data.europa.eu/en/trening/what-open-data` (visited on 01/20/2023).

[176] Centre for Data Ethics and Innovation. *Public attitudes to data and AI: Tracker survey.* en. 2022. URL: `https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey` (visited on 01/25/2023).

[177] Hannah Spiro and Holly Clarke. *Monitoring how public attitudes towards data and AI vary over time.* en. Mar. 2022. URL: `https://cdei.blog.gov.uk/2022/03/30/monitoring-how-public-attitudes-towards-data-and-ai-vary-over-time/` (visited on 01/25/2023).

[178] George W. Warren and Ragnar Lofstedt. "Risk communication and COVID-19 in Europe: lessons for future public health crises". In: *Journal of Risk Research* 25.10 (2022), pp. 1161–1175. DOI: `10.1080/13669877.2021.1947874`. eprint: `https://doi.org/10.1080/13669877.2021.1947874`. URL: `https://doi.org/10.1080/13669877.2021.1947874`.

[179] Morgan Meaker. “Coronavirus is putting Europe’s privacy protectors front and centre – and they’re coming up short”. en. In: *The Correspondent* (2020). URL: `https://thecorrespondent.com/718/coronavirus-is-putting-europes-privacy-protectors-front-and-centre-and-theyre-coming-up-short/94083803484-19ad6e37` (visited on 10/02/2020).

[180] Morgan Meaker. “Why is Covid-19 surveillance tech welcomed in some countries but rejected in others?” en. In: *The Correspondent* (2020). URL: `https://thecorrespondent.com/691/why-is-covid-19-surveillance-tech-welcomed-in-some-countries-but-rejected-in-others/90545833158-3800b6c3` (visited on 10/05/2020).

[181] European Centre for Disease Prevention and Control. *COVID-19: Latest news and reports.* en. 2023. URL: `https://www.ecdc.europa.eu/en/covid-19` (visited on 01/20/2023).

[182] Morgan Meaker. “From Japan to Brazil and South Africa: how countries’ ‘data cultures’ shape their response to coronavirus”. en. In: *The Correspondent* (2020). URL: `https://thecorrespondent.com/639/from-japan-to-brazil-and-south-africa-how-countries-data-cultures-shape-their-response-to-coronavirus/83731964382-c0c3bb35` (visited on 08/17/2020).

[183] Karola Klatt. “Corona apps: South Korea and the dark side of digital tracking”. en. In: *The Brussels Times* (2020). URL: `https://www.brusselstimes.com/108594/corona-apps-south-korea-and-the-dark-side-of-digital-tracking` (visited on 01/20/2023).

[184] Nicola Perra and Bruno Gonçalves. “Modeling and Predicting Human Infectious Diseases”. In: *Social Phenomena* (Apr. 2015), pp. 59–83. DOI: `10.1007/978-3-319-14011-7_4`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7123706/` (visited on 01/27/2023).

[185] Junwen Tao et al. “How to improve infectious disease prediction by integrating environmental data: an application of a novel ensemble analysis strategy to predict HFMD”. In: *Epidemiology and Infection* 149 (Jan. 2021), e34. ISSN: 0950-2688. DOI: `10.1017/S0950268821000091`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8060825/` (visited on 07/31/2022).

[186] Harshavardhan Achrekar et al. “Predicting Flu Trends using Twitter data”. In: *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS).* 2011, pp. 702–707. DOI: `10.1109/INFCOMW.2011.5928903`.

[187] Ali Alessa and Miad Faezipour. “Preliminary Flu Outbreak Prediction Using Twitter Posts Classification and Linear Regression With Historical Centers for Disease Control and Prevention Reports: Prediction Framework Study”. In: *JMIR Public Health and Surveillance* 5.2 (June 2019), e12383. ISSN: 2369-2960. DOI: `10.2196/12383`. URL:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6615001/ (visited on 09/26/2022).

[188] E. E. Rees et al. “Risk assessment strategies for early detection and prediction of infectious disease outbreaks associated with climate change”. eng. In: *Canada Communicable Disease Report = Releve Des Maladies Transmissibles Au Canada* 45.5 (May 2019), pp. 119–126. ISSN: 1188-4169. DOI: 10.14745/ccdr.v45i05a02.

[189] Rens van de Schoot et al. “Bayesian statistics and modelling”. en. In: *Nature Reviews Methods Primers* 1.1 (Jan. 2021), pp. 1–26. ISSN: 2662-8449. DOI: 10.1038/s43586-020-00001-2. URL: https://www.nature.com/articles/s43586-020-00001-2 (visited on 03/03/2023).

[190] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. “Probabilistic programming in Python using PyMC3”. In: *PeerJ Computer Science* 2 (Apr. 2016), e55. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.55. URL: https://peerj.com/articles/cs-55 (visited on 03/21/2019).

[191] Angela Noufaily et al. “An improved algorithm for outbreak detection in multiple surveillance systems”. In: *Statistics in Medicine* 32.7 (2013), pp. 1206–1222. ISSN: 02776715. DOI: 10.1002/sim.5595.

[192] Maximilian Gertler et al. “Outbreak of following river flooding in the city of Halle (Saale), Germany, August 2013”. In: *BMC Infectious Diseases* 15.1 (2015), pp. 1–10. ISSN: 14712334. DOI: 10.1186/s12879-015-0807-1.

[193] Maëlle Salmon et al. “A system for automated outbreak detection of communicable diseases in Germany”. eng. In: *Eurosurveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 21.13 (2016). ISSN: 1560-7917. DOI: 10.2807/1560-7917.ES.2016.21.13.30180.

[194] Martin Kulldorff. “A spatial scan statistic”. In: *Communications in Statistics - Theory and Methods* 26.6 (Jan. 1997), pp. 1481–1496. ISSN: 0361-0926. DOI: 10.1080/03610929708831995. URL: https://doi.org/10.1080/03610929708831995 (visited on 09/27/2018).

[195] Martin Kulldorff et al. “A space-time permutation scan statistic for disease outbreak detection”. In: *PLoS Medicine* 2.3 (2005), pp. 0216–0224. ISSN: 15491277. DOI: 10.1371/journal.pmed.0020059.

[196] Sebastian Meyer, Leonhard Held, and Michael Höhle. “Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance”. en-US. In: *Journal of Statistical Software* (2017). DOI: 10.18637/jss.v077.i11. URL: https://www.jstatsoft.org/article/view/v077i11 (visited on 09/27/2018).

[197] Cathy W. S. Chen, Khemmanant Khamthong, and Sangyeol Lee. “Markov switching integer-valued generalized auto-regressive conditional heteroscedastic models for dengue counts”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68.4 (2019), pp. 963–983. DOI: 10.1111/rssc.12344. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssc.12344. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12344.

[198] Yingcun Xia, Ottar N. Bjørnstad, and Bryan T. Grenfell. “Measles Metapopulation Dynamics: A Gravity Model for Epidemiological Coupling and Dynamics.” In: *The American Naturalist* 164.2 (Aug. 2004), pp. 267–281. ISSN: 0003-0147. DOI: 10.1086/422341. URL: https://www.journals.uchicago.edu/doi/10.1086/422341 (visited on 11/29/2018).

[199] Leonhard Held and Sebastian Meyer. “Forecasting Based on Surveillance Data”. In: *arXiv:1809.03735 [stat]* (Sept. 2018). arXiv: 1809.03735. URL: http://arxiv.org/abs/1809.03735 (visited on 03/27/2019).

[200] P McCullagh and J.A. Nelder. *Generalized Linear Models*. en. 2nd. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1989. ISBN: 978-0-412-31760-6.

[201] J-H Lee et al. “Analysis of overdispersed count data: application to the Human Papillomavirus Infection in Men (HIM) Study”. In: *Epidemiology & Infection* 140.6 (2012), pp. 1087–1094.

[202] John Gurland. “Some Applications of the Negative Binomial and Other Contagious Distributions”. In: *American Journal of Public Health and the Nations Health* 49.10 (Oct. 1959), pp. 1388–1399. ISSN: 0002-9572. DOI: 10.2105/AJPH.49.10.1388. URL: https://ajph.aphapublications.org/doi/abs/10.2105/AJPH.49.10.1388.

[203] Sylvain Coly et al. “Distributions to model overdispersed count data”. In: *Journal de la Societe Française de Statistique* 157.2 (2016), pp. 39–63. URL: https://hal.archives-ouvertes.fr/hal-01606783.

[204] Andrew Gelman. “Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)”. In: *Bayesian Analysis* 1.3 (Sept. 2006), pp. 515–534. ISSN: 1936-0975, 1931-6690. DOI: 10.1214/06-BA117A. URL: https://projecteuclid.org/euclid.ba/1340371048 (visited on 03/21/2019).

[205] Matthew D. Hoffman and Andrew Gelman. “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo”. In: *arXiv:1111.4246 [cs, stat]* (Nov. 2011). arXiv: 1111.4246. URL: http://arxiv.org/abs/1111.4246 (visited on 03/21/2019).

[206] Andrew Gelman and Donald B. Rubin. “Inference from Iterative Simulation Using Multiple Sequences”. In: *Statistical Science* 7.4 (1992), pp. 457–472. ISSN: 0883-4237. URL: https://www.jstor.org/stable/2246093.

[207] Carl De Boor. “On calculating with B-splines”. In: *Journal of Approximation theory* 6.1 (1972), pp. 50–62.

[208] World Health Organization and Food and Agriculture Organization of the United Nations and World Organisation for Animal Health. *The global view of campylobacteriosis: report of an expert consultation, Utrecht, Netherlands, 9-11 July 2012*. en. World Health Organization, 2013. ISBN: 9789241564601. URL: `https://apps.who.int/iris/handle/10665/80751`.

[209] Umesh D Parashar, E Anthony S Nelson, and Gagandeep Kang. “Diagnosis, management, and prevention of rotavirus gastroenteritis in children”. In: *BMJ (Clinical research ed.)* 347 (Dec. 2013), f7204. ISSN: 0959-8138. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5776699/`.

[210] Allen C. Steere et al. “Lyme borreliosis”. In: *Nature reviews. Disease primers* 2 (Dec. 2016), p. 16090. ISSN: 2056-676X. DOI: `10.1038/nrdp.2016.90`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5539539/`.

[211] Alois Stutzer and Bruno S Frey. “Commuting and Life Satisfaction in Germany”. In: *Informationen zur Raumentwicklung* (2007).

[212] Michael C Burda and Mark Weder. *The Economics of German Unification after Twenty-five Years: Lessons for Korea*. en. SFB 649 Discussion Papers SFB649DP2017-009. Sonderforschungsbereich 649, Humboldt University, Berlin, Germany, 2017, p. 30. URL: `https://ideas.repec.org/p/hum/wpaper/sfb649dp2017-009.html`.

[213] Marta Zawilska-Florczuk and Artur Ciechanowicz. *One country, two societies? Germany twenty years after reunification*. en. Tech. rep. Centre for Eastern Studies, Feb. 2011. URL: `https://www.osw.waw.pl/en/publikacje/osw-studies/2011-02-15/one-country-two-societies-germany-twenty-years-after-reunification` (visited on 10/15/2018).

[214] Sumio Watanabe. “A Widely Applicable Bayesian Information Criterion”. In: *J. Mach. Learn. Res.* 14.1 (Mar. 2013), pp. 867–897. ISSN: 1532-4435. URL: `http://dl.acm.org/citation.cfm?id=2502581.2502609`.

[215] Andrew Gelman, Jessica Hwang, and Aki Vehtari. “Understanding predictive information criteria for Bayesian models”. In: *Statistics and Computing* 24.6 (Nov. 2014), pp. 997–1016. ISSN: 1573-1375. DOI: `10.1007/s11222-013-9416-2`. URL: `https://doi.org/10.1007/s11222-013-9416-2`.

[216] A. Philip Dawid and Paola Sebastiani. “Coherent dispersion criteria for optimal experimental design”. en. In: *The Annals of Statistics* 27.1 (Mar. 1999), pp. 65–81. ISSN: 0090-5364, 2168-8966. DOI: `10.1214/aos/1018031101`. URL: `https://projecteuclid.org/euclid.aos/1018031101` (visited on 01/29/2019).

[217] Sebastian Meyer, Leonhard Held, and Michael Höhle. “hhh4: Endemic-epidemic modeling of areal count time series”. In: 1 (2016).

[218] Jay M. Ver Hoef and Peter L. Boveng. “Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?” eng. In: *Ecology* 88.11 (Nov. 2007), pp. 2766–2772. ISSN: 0012-9658.

[219] Andrew B. Lawson. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology.* 3rd. Chapman & Hall/CRC Interdisciplinary Statistics. New York: Chapman and Hall/CRC, 2018. ISBN: 9781351271769.

[220] Sudipto Banerjee et al. *Handbook of Spatial Epidemiology.* 1st. Chapman & Hall/CRC handbooks of modern statistical methods. New York: Chapman and Hall/CRC, 2016. ISBN: 9781482253016.

[221] Juliane Manitz et al. “Origin Detection During Food-borne Disease Outbreaks – A Case Study of the 2011 EHEC/HUS Outbreak in Germany”. English. In: *PLOS Currents Outbreaks* (Apr. 2014). ISSN: 2157-3999. DOI: 10.1371/currents.outbreaks.f3fdeb08c5b9de7c09ed9cbcef5f01f2. URL: index.html%3Fp=36515.html (visited on 03/13/2019).

[222] Sebastian Meyer, Leonhard Held, et al. “Power-law models for infectious disease spread”. In: *The Annals of Applied Statistics* 8.3 (2014), pp. 1612–1639.

[223] D. J. Watson. “Comparative Physiological Studies on the Growth of Field Crops: I. Variation in Net Assimilation Rate and Leaf Area between Species and Varieties, and within and between Years”. In: *Annals of Botany* 11.1 (Jan. 1947), pp. 41–76. ISSN: 0305-7364. DOI: 10.1093/oxfordjournals.aob.a083148. URL: https://academic.oup.com/aob/article/11/1/41/159526 (visited on 06/20/2019).

[224] J. M. Chen and T. A. Black. “Defining leaf area index for non-flat leaves”. In: *Plant, Cell & Environment* 15.4 (1992), pp. 421–429. DOI: 10.1111/j.1365-3040.1992.tb00992.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-3040.1992.tb00992.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3040.1992.tb00992.x.

[225] J. Montheith and M. Unsworth. *Principles of Environmental Physics.* Academic Press, San Diego, CA, USA, 2007.

[226] Guangjian Yan et al. “Review of indirect optical measurements of leaf area index: Recent advances, challenges, and perspectives”. In: *Agricultural and Forest Meteorology* 265 (Feb. 2019), pp. 390–411. ISSN: 0168-1923. DOI: 10.1016/j.agrformet.2018.11.033. URL: http://www.sciencedirect.com/science/article/pii/S0168192318303873 (visited on 06/20/2019).

[227] Sidney Cox. “Information technology: The global key to precision agriculture and sustainability”. In: *Computers and Electronics in Agriculture* 36 (Nov. 2002), pp. 93–111. DOI: 10.1016/S0168-1699(02)00095-9.

[228] David J. Mulla. “Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps”. In: *Biosystems Engineering* 114.4 (2013). Special Issue: Sensing Technologies for Sustainable Agriculture, pp. 358–371. ISSN: 1537-5110. DOI: https://doi.org/10.1016/j.biosystemseng.2012.08.009. URL: http://www.sciencedirect.com/science/article/pii/S1537511012001419.

[229] K.H. Kjaer, K.K. Petersen, and M. Bertelsen. “Protective rain shields alter leaf microclimate and photosynthesis in organic apple production”. In: *Acta Horticulturae* 1134 (May 2016), pp. 317–326. ISSN: 0567-7572, 2406-6168. DOI: 10.17660/ActaHortic.2016.1134.42. URL: https://www.actahort.org/books/1134/1134_42.htm (visited on 05/24/2021).

[230] Mairaj Din et al. “Evaluating Hyperspectral Vegetation Indices for Leaf Area Index Estimation of Oryza sativa L. at Diverse Phenological Stages”. In: *Frontiers in Plant Science* 8 (2017), p. 820. ISSN: 1664-462X. DOI: 10.3389/fpls.2017.00820. URL: https://www.frontiersin.org/article/10.3389/fpls.2017.00820.

[231] Gregory L. Britten et al. “Evaluating the Benefits of Bayesian Hierarchical Methods for Analyzing Heterogeneous Environmental Datasets: A Case Study of Marine Organic Carbon Fluxes”. In: *Frontiers in Environmental Science* 9 (2021), p. 28. ISSN: 2296-665X. DOI: 10.3389/fenvs.2021.491636. URL: https://www.frontiersin.org/article/10.3389/fenvs.2021.491636.

[232] Cornelius Senf et al. “A Bayesian hierarchical model for estimating spatial and temporal variation in vegetation phenology from Landsat time series”. In: *Remote Sensing of Environment* 194 (2017), pp. 155–160. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2017.03.020. URL: https://www.sciencedirect.com/science/article/pii/S0034425717301232.

[233] Chad Babcock, Andrew O. Finley, and Nathaniel Looker. “A Bayesian model to estimate land surface phenology parameters with harmonized Landsat 8 and Sentinel-2 images”. In: *Remote Sensing of Environment* 261 (2021), p. 112471. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2021.112471. URL: https://www.sciencedirect.com/science/article/pii/S0034425721001899.

[234] Tong Qiu et al. “Understanding the continuous phenological development at daily time step with a Bayesian hierarchical space-time model: impacts of climate change and extreme weather events”. In: *Remote Sensing of Environment* 247 (2020), p. 111956. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2020.111956. URL: https://www.sciencedirect.com/science/article/pii/S0034425720303266.

[235] Bijan Seyednasrollah et al. “Leaf phenology paradox: Why warming matters most where it is already warm”. In: *Remote Sensing of Environment* 209 (May 2018), pp. 446–455. DOI: 10.1016/j.rse.2018.02.059.

[236] Adam M. Wilson et al. “Scaling up: linking field data and remote sensing with a hierarchical model”. In: *International Journal of Geographical Information Science* 25.3 (2011), pp. 509–521. DOI: c. eprint: https://doi.org/10.1080/13658816.2010.522779. URL: https://doi.org/10.1080/13658816.2010.522779.

[237] Gordon Pipa et al. “Mapping of Visual Receptive Fields by Tomographic Reconstruction”. In: *Neural computation* 24.10 (Oct. 2012), pp. 2543–2578. ISSN: 0899-7667. DOI: 10.1162/NECO_a_00334. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3972919/ (visited on 03/04/2019).

[238] Nathalie J. J. Bréda. “Ground-based measurements of leaf area index: a review of methods, instruments and current controversies”. In: *Journal of Experimental Botany* 54.392 (Nov. 2003), pp. 2403–2417. ISSN: 0022-0957. DOI: 10.1093/jxb/erg263. URL: https://academic.oup.com/jxb/article/54/392/2403/621920 (visited on 06/20/2019).

[239] John K. Schueller. “A review and integrating analysis of Spatially-Variable Control of crop production”. In: *Fertilizer research* 33.1 (Oct. 1992), pp. 1–34. ISSN: 1573-0867. DOI: 10.1007/BF01058007. URL: https://doi.org/10.1007/BF01058007.

[240] Laurent Kergoat et al. “Impact of doubled CO2 on global-scale leaf area index and evapotranspiration: Conflicting stomatal conductance and LAI responses”. In: *Journal of Geophysical Research: Atmospheres* 107.D24 (2002), ACL 30-1-ACL 30–16. DOI: 10.1029/2001JD001245. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2001JD001245. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001JD001245.

[241] H. Yan et al. “Global estimation of evapotranspiration using a leaf area index-based surface energy and water balance model”. In: *Remote Sensing of Environment* 124 (2012), pp. 581–595. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2012.06.004. URL: http://www.sciencedirect.com/science/article/pii/S0034425712002404.

[242] Simic Milas Anita, Fernandes Richard, and Shusen Wang. “Assessing the Impact of Leaf Area Index on Evapotranspiration and Groundwater Recharge across a Shallow Water Region for Diverse Land Cover and Soil Properties”. In: *Journal of Water Resource and Hydraulic Engineering* 3 (Dec. 2014), pp. 60–73.

[243] Gregory P. Asner, Jonathan M. O. Scurlock, and Jeffrey A. Hicke. “Global synthesis of leaf area index observations: implications for ecological and remote sensing studies”. In: *Global Ecology and Biogeography* 12.3 (2003), pp. 191–205. ISSN: 1466-8238. DOI: 10.1046/j.1466-822X.2003.00026.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1466-822X.2003.00026.x (visited on 06/20/2019).

[244] N.H Broge and J.V Mortensen. "Deriving green crop area index and canopy chlorophyll density of winter wheat from spectral reflectance data". In: *Remote Sensing of Environment* 81.1 (2002), pp. 45–57. ISSN: 0034-4257. DOI: https://doi.org/10.1016/S0034-4257(01)00332-7. URL: https://www.sciencedirect.com/science/article/pii/S0034425701003327.

[245] M. S. Moran, S. J. Maas, and P. J. Pinter Jr. "Combining remote sensing and modeling for estimating surface evaporation and biomass production". In: *Remote Sensing Reviews* 12.3-4 (1995), pp. 335–353. DOI: 10.1080/02757259509532290. eprint: https://doi.org/10.1080/02757259509532290. URL: https://doi.org/10.1080/02757259509532290.

[246] Zhiqiang Gao, Wei Gao, and James Slusser. "The response of leaf area index to climate change during 1981-2000 in China". In: *Remote Sensing and Modeling of Ecosystems for Sustainability II*. Vol. 5884. International Society for Optics and Photonics, Sept. 2005, 58840S. DOI: 10.1117/12.612929. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5884/58840S/The-response-of-leaf-area-index-to-climate-change-during/10.1117/12.612929.short (visited on 09/24/2020).

[247] Anthony Manea and Michelle R. Leishman. "Leaf Area Index Drives Soil Water Availability and Extreme Drought-Related Mortality under Elevated CO2 in a Temperate Grassland Model System". In: *PLoS ONE* 9.3 (Mar. 2014). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0091046. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3954624/ (visited on 11/06/2020).

[248] Inge Jonckheere et al. "Review of methods for in situ leaf area index determination: Part I. Theories, sensors and hemispherical photography". In: *Agricultural and Forest Meteorology* 121.1 (2004), pp. 19–35. ISSN: 0168-1923. DOI: https://doi.org/10.1016/j.agrformet.2003.08.027. URL: https://www.sciencedirect.com/science/article/pii/S0168192303001643.

[249] Li He et al. "Comparing methods for estimating leaf area index by multi-angular remote sensing in winter wheat". In: *Scientific Reports* 10 (Aug. 2020). DOI: 10.1038/s41598-020-70951-w.

[250] Ke Liu et al. "Estimating the crop leaf area index using hyperspectral remote sensing". In: *Journal of Integrative Agriculture* 15.2 (2016), pp. 475–491. ISSN: 2095-3119. DOI: https://doi.org/10.1016/S2095-3119(15)61073-5. URL: https://www.sciencedirect.com/science/article/pii/S2095311915610735.

[251] A Huete et al. "Overview of the radiometric and biophysical performance of the MODIS vegetation indices". In: *Remote Sensing of Environment* 83.1 (2002). The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring, pp. 195–213. ISSN: 0034-4257.

DOI: https://doi.org/10.1016/S0034-4257(02)00096-2. URL: https://www.sciencedirect.com/science/article/pii/S0034425702000962.

[252] W.A. Dorigo et al. “A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling”. In: *International Journal of Applied Earth Observation and Geoinformation* 9.2 (2007). Advances in airborne electromagnetics and remote sensing of agro-ecosystems, pp. 165–193. ISSN: 0303-2434. DOI: https://doi.org/10.1016/j.jag.2006.05.003. URL: https://www.sciencedirect.com/science/article/pii/S0303243406000201.

[253] Yves M. Govaerts et al. “Designing optimal spectral indices: A feasibility and proof of concept study”. In: *International Journal of Remote Sensing* 20.9 (1999), pp. 1853–1873. DOI: 10.1080/014311699212524. eprint: https://doi.org/10.1080/014311699212524. URL: https://doi.org/10.1080/014311699212524.

[254] Anthony Nguy-Robertson et al. “Estimating green LAI in four crops: Potential of determining optimal spectral bands for a universal algorithm”. In: *Agricultural and Forest Meteorology* s 192–193 (July 2014), pp. 140–148. DOI: 10.1016/j.agrformet.2014.03.004.

[255] Lydia Serrano, Iolanda Filella, and Josep Penuelas. “Remote Sensing of Biomass and Yield of Winter Wheat under Different Nitrogen Supplies”. In: *Crop Science* 40 (May 2000), pp. 723–731. DOI: 10.2135/cropsci2000.403723x.

[256] Yaseen T. Mustafa, Valentyn A. Tolpekin, and Alfred Stein. “Improvement of Spatio-temporal Growth Estimates in Heterogeneous Forests Using Gaussian Bayesian Networks”. In: *IEEE Transactions on Geoscience and Remote Sensing* 52.8 (2014), pp. 4980–4991. DOI: 10.1109/TGRS.2013.2286219.

[257] Huaan Jin et al. “Spatially and Temporally Continuous Leaf Area Index Mapping for Crops through Assimilation of Multi-resolution Satellite Data”. In: *Remote Sensing* 11.21 (2019). ISSN: 2072-4292. DOI: 10.3390/rs11212517. URL: https://www.mdpi.com/2072-4292/11/21/2517.

[258] Jiayi Ji et al. “Multiscale leaf area index assimilation for Moso bamboo forest based on Sentinel-2 and MODIS data”. In: *International Journal of Applied Earth Observation and Geoinformation* 104 (2021), p. 102519. ISSN: 0303-2434. DOI: https://doi.org/10.1016/j.jag.2021.102519. URL: https://www.sciencedirect.com/science/article/pii/S0303243421002269.

[259] Juanjuan Zhang et al. “Leaf area index estimation model for UAV image hyperspectral data based on wavelength variable selection and machine learning methods”. In: *Plant Methods* 17.1 (May 3, 2021), p. 49. ISSN: 1746-4811. DOI: 10.1186/s13007-021-00750-5. URL: https://doi.org/10.1186/s13007-021-00750-5 (visited on 05/06/2021).

[260] Liang Wan et al. “Unmanned aerial vehicle-based field phenotyping of crop biomass using growth traits retrieved from PROSAIL model”. In: *Computers and Electronics in Agriculture* 187 (2021), p. 106304. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2021.106304. URL: https://www.sciencedirect.com/science/article/pii/S0168169921003215.

[261] Bo Sun et al. “Retrieval of rapeseed leaf area index using the PROSAIL model with canopy coverage derived from UAV images as a correction parameter”. In: *International Journal of Applied Earth Observation and Geoinformation* 102 (2021), p. 102373. ISSN: 0303-2434. DOI: https://doi.org/10.1016/j.jag.2021.102373. URL: https://www.sciencedirect.com/science/article/pii/S0303243421000805.

[262] Bastian Siegmann and Thomas Jarmer. “Comparison of different regression models and validation techniques for the assessment of wheat leaf area index from hyperspectral data”. In: *International Journal of Remote Sensing* 36.18 (2015), pp. 4519–4534. DOI: 10.1080/01431161.2015.1084438. eprint: https://doi.org/10.1080/01431161.2015.1084438. URL: https://doi.org/10.1080/01431161.2015.1084438.

[263] Luqi Xing et al. “Assimilating Multiresolution Leaf Area Index of Moso Bamboo Forest from MODIS Time Series Data Based on a Hierarchical Bayesian Network Algorithm”. In: *Remote Sensing* 11.1 (2019). ISSN: 2072-4292. DOI: 10.3390/rs11010056. URL: https://www.mdpi.com/2072-4292/11/1/56.

[264] Kusum J. Naithani et al. “Spatial Distribution of Tree Species Governs the Spatio-Temporal Interaction of Leaf Area Index and Soil Moisture across a Forested Landscape”. In: *PLOS ONE* 8.3 (Mar. 2013), pp. 1–12. DOI: 10.1371/journal.pone.0058704. URL: https://doi.org/10.1371/journal.pone.0058704.

[265] Daniel Schraik et al. “Bayesian inversion of a forest reflectance model using Sentinel-2 and Landsat 8 satellite images”. In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 233 (2019), pp. 1–12. ISSN: 0022-4073. DOI: https://doi.org/10.1016/j.jqsrt.2019.05.013. URL: https://www.sciencedirect.com/science/article/pii/S002240731930175X.

[266] X.Q. Xu et al. “Inversion of rice canopy chlorophyll content and leaf area index based on coupling of radiative transfer and Bayesian network models”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 150 (2019), pp. 185–196. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2019.02.013. URL: https://www.sciencedirect.com/science/article/pii/S0924271619300450.

[267] Rogério P. Soratto et al. “Agronomic optimal plant density for semiupright cowpea as a second crop in southeastern Brazil”. In: *Crop Science* 60.5 (2020), pp. 2695–2708. DOI: https://doi.org/10.1002/csc2.20232. eprint: https://acsess.onlinelibrary.wiley.com/doi/pdf/10.1002/csc2.20232.

URL: https://acsess.onlinelibrary.wiley.com/doi/abs/10.1002/csc2.20232.

[268] Yonghua Qu et al. "A Bayesian network algorithm for retrieving the characterization of land surface vegetation". In: *Remote Sensing of Environment* 112.3 (2008), pp. 613–622. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2007.03.031. URL: https://www.sciencedirect.com/science/article/pii/S0034425707003896.

[269] Carl de Boor. *A Practical Guide to Splines*. Applied Mathematical Sciences. New York: Springer-Verlag, 1978. ISBN: 978-0-387-95366-3. URL: https://www.springer.com/de/book/9780387953663 (visited on 03/04/2019).

[270] Tegoeh Tjahjowidodo, VT Dung, and ML Han. "A fast non-uniform knots placement method for B-spline fitting". In: *2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE. 2015, pp. 1490–1495.

[271] Lluís Jordi Ferrer Arnau et al. "Efficient cubic spline interpolation implemented with FIR filters". In: *International Journal of Computer Information Systems and Industrial Management Applications* 105 (2013), pp. 98–105. ISSN: 2150-7988.

[272] Andrew Gelman et al. *Bayesian data analysis*. CRC press, 2013.

[273] P McCullagh and John A Nelder. *Generalized Linear Models*. Vol. 37. CRC Press, 1989.

[274] Aki Vehtari et al. "Pareto Smoothed Importance Sampling". In: *arXiv:1507.02646 [stat]* (July 2, 2019). arXiv: 1507.02646. URL: http://arxiv.org/abs/1507.02646 (visited on 09/05/2020).

[275] Yan-Ping Cen and Janet F. Borman. "The Response of Bean Plants to UV-B Radiation Under Different Irradiances of Background Visible Light". In: *Journal of Experimental Botany* 41.11 (Nov. 1990), pp. 1489–1495. ISSN: 0022-0957. DOI: 10.1093/jxb/41.11.1489. eprint: https://academic.oup.com/jxb/article-pdf/41/11/1489/1249515/41-11-1489.pdf. URL: https://doi.org/10.1093/jxb/41.11.1489.

[276] Nieves Aparicio et al. "Spectral Vegetation Indices as Nondestructive Tools for Determining Durum Wheat Yield". In: *Agronomy Journal* 92.1 (2000), pp. 83–91. DOI: https://doi.org/10.2134/agronj2000.92183x. eprint: https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2134/agronj2000.92183x. URL: https://acsess.onlinelibrary.wiley.com/doi/abs/10.2134/agronj2000.92183x.

[277] Vivian Roca Schwendler Weber et al. "Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes". In: *Field Crops Research* 128 (2012). DOI: 10.1016/j.fcr.2011.12.016.

[278] Jianfeng Zhang et al. “Leaf Chlorophyll Content Estimation of Winter Wheat Based on Visible and Near-Infrared Sensors”. In: *Sensors (Basel, Switzerland)* 16 (Mar. 2016). DOI: 10.3390/s16040437.

[279] Daniel Sims and John Gamon. “Estimation of vegetation water content and photosynthetic tissue area from spectral reflectance: A comparison of indices based on liquid water and chlorophyll absorption features”. In: *Remote Sensing of Environment* 84 (Apr. 2003), pp. 526–537. DOI: 10.1016/S0034-4257(02)00151-7.

[280] Chao Wang et al. “Extraction of Sensitive Bands for Monitoring the Winter Wheat (Triticum aestivum) Growth Status and Yields Based on the Spectral Reflectance”. In: *PLOS ONE* 12.1 (Jan. 2017), pp. 1–16. DOI: 10.1371/journal.pone.0167679. URL: https://doi.org/10.1371/journal.pone.0167679.

[281] Judea Pearl. “Comment: understanding Simpson’s paradox”. In: *The American Statistician* 68.1 (2014), pp. 8–13.

[282] Ritika Srinet, Subrata Nandy, and N.R. Patel. “Estimating leaf area index and light extinction coefficient using Random Forest regression algorithm in a tropical moist deciduous forest, India”. In: *Ecological Informatics* 52 (2019), pp. 94–102. ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2019.05.008. URL: https://www.sciencedirect.com/science/article/pii/S1574954118303029.

[283] Rasmus Houborg and Matthew McCabe. “A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 135 (Jan. 2018), pp. 173–188. DOI: 10.1016/j.isprsjprs.2017.10.004.

[284] Tongtong Wang, Zhiqiang Xiao, and Zhigang Liu. “Performance Evaluation of Machine Learning Methods for Leaf Area Index Retrieval from Time-Series MODIS Reflectance Data”. In: *Sensors* 17.1 (2017). ISSN: 1424-8220. DOI: 10.3390/s17010081. URL: https://www.mdpi.com/1424-8220/17/1/81.

[285] Tomoaki Yamaguchi et al. “Feasibility of Combining Deep Learning and RGB Images Obtained by Unmanned Aerial Vehicle for Leaf Area Index Estimation in Rice”. In: *Remote Sensing* 13.1 (2021). ISSN: 2072-4292. DOI: 10.3390/rs13010084. URL: https://www.mdpi.com/2072-4292/13/1/84.

[286] Orly Enrique Apolo-Apolo et al. “A Mixed Data-Based Deep Neural Network to Estimate Leaf Area Index in Wheat Breeding Trials”. In: *Agronomy* 10.2 (2020). ISSN: 2073-4395. DOI: 10.3390/agronomy10020175. URL: https://www.mdpi.com/2073-4395/10/2/175.

[287] Aki Vehtari, Andrew Gelman, and Jonah Gabry. “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. In: *Statistics and Computing* 27.5 (Sept. 2017). arXiv: 1507.04544, pp. 1413–1432. ISSN: 0960-3174, 1573-1375. DOI: 10.1007/s11222-016-9696-4. URL: http://arxiv.org/abs/1507.04544 (visited on 09/05/2020).

[288] Judea Pearl et al. “Causal inference in statistics: An overview”. In: *Statistics surveys* 3 (2009), pp. 96–146.

[289] Andrew Richardson, Shane P. Duigan, and Graeme Berlyn. “An evaluation of noninvasive methods to estimate foliar chlorophyll content”. In: *New Phytologist* 153 (Jan. 2002), pp. 185–194. DOI: 10.1046/j.0028-646X.2001.00289.x.

[290] Liangyun Liu et al. “Estimating winter wheat plant water content using red edge parameters”. In: *International Journal of Remote Sensing* 25.17 (2004), pp. 3331–3342. DOI: 10.1080/01431160310001654365. eprint: https://doi.org/10.1080/01431160310001654365. URL: https://doi.org/10.1080/01431160310001654365.

[291] J.G.P.W. Clevers and Lammert Kooistra. “Using hyperspectral remote sensing data for retrieving canopy water content”. In: *WHISPERS ’09 - 1st Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing* (Sept. 2009), pp. 1–4. DOI: 10.1109/WHISPERS.2009.5289058.

[292] Yisong Cheng et al. “Spectral red edge parameters for winter wheat under different nitrogen support levels”. In: *Proceedings of SPIE - The International Society for Optical Engineering* (Aug. 2005). DOI: 10.1117/12.614759.

[293] Vali Rasooli Sharabiani, Noboru Noguchi, and Kazunobu Ishi. “Significant wavelengths for prediction of winter wheat growth status and grain yield using multivariate analysis”. In: *Engineering in Agriculture, Environment and Food* 7 (Feb. 2014), pp. 14–21. DOI: 10.1016/j.eaef.2013.12.003.

[294] Wenjiang Huang et al. “Application of red edge variables in winter wheat nutrition diagnosis”. In: *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium.* Vol. 6. 2004, 4052–4055 vol.6. DOI: 10.1109/IGARSS.2004.1370020.

[295] M.H. Rad et al. “Effects of Different Soil Moisture Regimes on Leaf Area Index, Specific Leaf Area and Water use Efficiency in Eucalyptus (Eucalyptus camaldulensis Dehnh) under Dry Climatic Conditions”. In: *Asian Journal of Plant Sciences* 10 (2011), pp. 294–300. DOI: 10.3923/ajps.2011.294.300.

[296] Daniel T. C. Cox et al. “Global variation in diurnal asymmetry in temperature, cloud cover, specific humidity and precipitation and its association with leaf area index”. In: *Global Change Biology* n/a.n/a (2020). ISSN: 1365-2486. DOI: 10.1111/gcb.15336. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15336 (visited on 11/07/2020).

[297] Stephen R. Hardwick et al. “The relationship between leaf area index and microclimate in tropical forest and oil palm plantation: Forest disturbance drives changes in microclimate”. In: *Agricultural and Forest Meteorology* 201 (Feb. 2015), pp. 187–195. ISSN: 0168-1923. DOI: 10.1016/j.agrformet.2014.11.010. URL: http:

//www.sciencedirect.com/science/article/pii/S0168192314002780 (visited on 11/07/2020).

[298] Lijun Su et al. "Simulation Models of Leaf Area Index and Yield for Cotton Grown with Different Soil Conditioners". In: *PLoS ONE* 10.11 (Nov. 2015). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0141835. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633134/ (visited on 11/07/2020).

[299] John W. Tukey. "The Future of Data Analysis". EN. In: *The Annals of Mathematical Statistics* 33.1 (Mar. 1962), pp. 1–67. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177704711. URL: https://projecteuclid.org/euclid.aoms/1177704711 (visited on 07/15/2019).

[300] EU Commision. *The EU General Data Protection Regulation (GDPR).* en-GB. 2018. URL: https://eugdpr.org/the-regulation/.

[301] Bryce Goodman and Seth Flaxman. "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"". en. In: *AI Magazine* 38.3 (2017), pp. 50–57. ISSN: 2371-9621. DOI: 10.1609/aimag.v38i3.2741. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/2741 (visited on 01/06/2022).

[302] Nicholas Vollmer. *Article 22 EU General Data Protection Regulation (EU-GDPR).* en. text. July 2021. URL: https://www.privacy-regulation.eu/en/article-22-automated-individual-decision-making-including-profiling-GDPR.htm (visited on 01/06/2022).

[303] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR". In: *Harvard journal of law & technology* 31 (Apr. 2018), pp. 841–887. DOI: 10.2139/ssrn.3063289.

[304] *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.* en. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206 (visited on 01/06/2022).

[305] High-Level Expert Group on Artificial Intelligence EU Commision. *Ethics Guidelines for Trustworthy AI.* Tech. rep. 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[306] Mark MacCarthy and Kenneth Propp. *Machines learn that Brussels writes the rules: The EU's new AI regulation.* en-US. May 2021. URL: https://www.brookings.edu/blog/techtank/2021/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/ (visited on 01/06/2022).

[307] Melissa Heikkilä. "The EU wants to put companies on the hook for harmful AI". en. In: *MIT Technology Review* (2022). URL: https://www.technologyreview.com/2022/10/01/1060539/eu-tech-policy-harmful-ai-liability/ (visited on 10/31/2022).

[308] European Commission. *Liability Rules for Artificial Intelligence*. en. 2022. URL: https://ec.europa.eu/info/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en (visited on 10/31/2022).

[309] Office of Science and Technology Policy. *Blueprint for an AI Bill of Rights - OSTP*. en-US. 2022. URL: https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (visited on 10/31/2022).

[310] Sigal Samuel. *There's something missing from the White House's AI ethics blueprint*. en. Oct. 2022. URL: https://www.vox.com/future-perfect/23387228/ai-bill-of-rights-white-house-artificial-intelligence-bias (visited on 10/31/2022).

[311] Cyberspace Administration of China. *Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022*. en. 2022. URL: https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/ (visited on 10/31/2022).

[312] Jennifer Conrad. "China Is About to Regulate AI—and the World Is Watching". en-US. In: *Wired* (2022). ISSN: 1059-1028. URL: https://www.wired.com/story/china-regulate-ai-world-watching/ (visited on 10/31/2022).

[313] Sapni G K and Mihir Mahajan. "Understanding China's Draft Algorithm Regulations". en-US. In: *The Diplomat* (2021). URL: https://thediplomat.com/2021/09/understanding-chinas-draft-algorithm-regulations/ (visited on 11/18/2022).

[314] Timnit Gebru et al. *Datasheets for Datasets*. 2018. DOI: 10.48550/ARXIV.1803.09010. URL: https://arxiv.org/abs/1803.09010.

[315] Eunsol Choi et al. "QuAC: Question Answering in Context". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2174–2184. DOI: 10.18653/v1/D18-1241. URL: https://aclanthology.org/D18-1241.

[316] Ismaïla Seck et al. *Baselines and a datasheet for the Cerema AWP dataset*. 2018. DOI: 10.48550/ARXIV.1806.04016. URL: https://arxiv.org/abs/1806.04016.

[317] Yang Trista Cao and Hal Daumé III. "Toward Gender-Inclusive Coreference Resolution". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 4568–4595. DOI: 10.18653/v1/2020.acl-main.418. URL: https://aclanthology.org/2020.acl-main.418.

[318] OHDSI Observational Health Data Sciences and Informatics. *OMOP Common Data Model – OHDSI*. en-US. 2020. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ (visited on 05/22/2020).

[319] Machine Learning and Market for Intelligence Conference 2016. *Geoffrey Hinton: On Radiology*. en. Toronto, CA, Nov. 2016. URL: https://www.youtube.com/watch?v=2HMPRXstSvQ (visited on 11/19/2022).

[320] Gary Smith and Jeffrey Funk. *AI has a long way to go before doctors can trust it with your life*. en. June 2021. URL: https://qz.com/2016153/ai-promised-to-revolutionize-radiology-but-so-far-its-failing/ (visited on 11/19/2022).