Geographic and Social Space in Latent Factor Models – Four Essays

Dissertationsarbeit

zur Erlangung des akademischen Grades Dr. rer. pol.

an der Wirtschaftswissenschaftlichen Fakultät der Universität Passau

Vorgelegt von

Angelika Schmid

Veröffentlichung der Dissertationsarbeit

zur Erlangung des akademischen Grades Dr. rer. pol.

an der Wirtschaftswissenschaftlichen Fakultät der Universität Passau

Dissertationsort: Passau

Tag der Disputation: 29. April 2019

1. Prüfer: Prof. Dr. Harry Haupt,

Lehrstuhl für Statistik und Data Analytics

2. Prüfer: Prof. Dr. Andreas König,

Lehrstuhl für Strategisches Management, Innovation und Entrepreneurship

Danksagung

Mein Dank gilt zunächst Herrn Prof. Dr. Harry Haupt, der mir die Chance eröffnet hat, mich einer Promotion im Bereich Statistik zu widmen, und für die Energie, die er in unsere gemeinsame Arbeit und meine Doktorarbeit investiert hat. Herrn Prof. Dr. König danke ich für das Zweitgutachten zu meiner Arbeit und für wertvolle Gedanken zu ihrer Verbesserung. Mein Dank gilt weiterhin meinen Koautoren und in dieser Funktion wieder Herrn Prof. Dr. Harry Haupt, Svenia Behm und Dr. Jörg Scheffer. Prof. Dr. Sven Apel und Thomas Bock sind wesentlich an einem der vier Artikel beteiligt, und mittlerweile meine Koautoren im Bereich Netzwerkanalyse und Open-Source Software Entwicklung mit Versionskontrolldaten.
Ich danke Svenia Behm, Dr. Joachim Schnurbus, Dr. Markus Fritsch und Dr. Sandra Huber für die gemeinsame Zeit als Team am Lehrstuhl. Ich bedanke mich bei meinen KollegInnen an der Wirtschaftswissenschaftlichen Fakultät für den wissenschaftlichen Austausch, mentale Unterstützung und das soziale Leben rund um die Promotion in Passau. Die Bearbeitungszeit und das Themengebiet dieser Doktorarbeit fällt zudem teils auch in meine Zeit an der Fakultät für Informatik. Dort durfte ich mit Barbara Eckl, Tina Schuh, Thomas Bock, Gustavo do Vale, Prof. Dr. Janet Siegmund, Prof. Dr. Sven Apel, sowie weiteren Promovierenden und Studierenden des Lehrstuhls für Software Engineering I an verschiedenen Projekten arbeiten und sehr viel lernen – vielen Dank auch dafür!

Thesis Abstract

Geography, social context, time, and cultural mindset are four (out of many) cornerstones of human interaction. When building statistical models, their consideration is vital: They all cause dependency between individual observations, violating assumptions of independence and exchangeability. While this can be problematic and inhibit the unbiased inference of parameters, it can also be a fruitful source of insights and enhance prediction performance.

One class of models that serves to manage or profit from the presence of dependence is the class of *latent variable models*. This class of models assumes that the presence of non-explicit, unobserved causes of continuous or discrete nature can explain the observed correlations. Latent variable models explicitly take account of dependency, for example, by modeling an unobserved local source of pollution as a continuous spatial variable. Through their widespread use for information filtering, link prediction, and statistical inference, latent variable models have developed an essential impact on our daily life and the way we consume information.

The four articles in this thesis shed light on assumptions, usage, and potential drawbacks of latent variable models in various contexts that involve geographic and interaction data. We model unobserved sources of pollution in geophysical data, explore individual taste and mindsets in cross-cultural contexts, and predict the evolution of social relationships in software development projects. This combination of various perspectives contributes to the interdisciplinary exchange of methodological knowledge on the modeling of dependent data.

Outline

Notation

Example	Description and Designation
a	Lowercase italic letters: scalars
$\mathbf{a}$	Lowercase boldface letters: Column vectors
$\mathbf{A} = \mathbf{A}^{m \times n}$	Capital boldface letters: $(m \times n)$-dimensional matrices
$\underline{\mathbf{A}} = \underline{\mathbf{A}}^{m \times n \times c}$	Underlined capital boldface letters: Tensors (here: third-order)
$\mathbf{A}_{(\cdot i)} = \mathbf{a}_i$	The i-th column of matrix $\mathbf{A}$
$\mathbf{A}_{(i \cdot)}$	The i-th row of matrix $\mathbf{A}$
$\mathbf{a}^T$	Row vector (transposed)
$\mathbf{A}^T$	Transpose of a matrix
$\|\|\mathbf{a}\|\|, \|\|\mathbf{A}\|\|$	Length/norm of a vector or matrix
$\|\|\mathbf{A}\|\|_2$	Spectral norm of a matrix
$\|\|\mathbf{A}\|\|_F$	Frobenius norm of a matrix
$(\mathbf{a}, \mathbf{b})$	Inner product
$\boldsymbol{\epsilon}$	Bold Greek letter $\boldsymbol{\epsilon}$: Vector of idiosyncratic errors
$\boldsymbol{\mu}^g$	Objects with superscript g: Group-specific processes or parameters
$\otimes$	Kronecker product
$E()$	Expected value
`R`	The `R`-project, open-source statistical software
`igraph`	Type-set words in Software context: `R` packages or objects
`plot()`	Type-set words with parentheses: `R` functions

List of Abbreviations

Acronym	Designation
AUC	Area Under Curve
CFI	Comparative Fit Index
DWLS	Diagonally Weighted Least Squares
HOSVD	High Order Singular Value Decomposition
LBSNs	Location Based Social Networks
LOOCV	Leave-One-Out Cross Validation
MCMC	Monte Carlo Markov Chain
PCA	Principal Component Analysis
PMF	Probabilistic Matrix Factorization
POI	Point Of Interest
RMSE	Root Mean Squared Error
RMSEA	Root Mean Squared Error of Approximation
ROC	Receiver Operating Characteristic
SVD	Singular Value Decomposition

List of Figures

List of Tables

Thesis Summary and Core Literature

A Introduction

In data analytics, there are often strong hypotheses on why these correlations among random variables arise, but the actual *cause* of the correlation cannot be observed. For example, the correlation between survey questions about the church and faith is related to the phenomenon of religious faith. Religious faith is not measurable, yet it manifests in correlation among different people's answers to a variety of survey questions. If faith could be modeled reliably via these observed responses, this measurement could be compared between groups, monitored over time, or analyzed with regards to its effects on other areas of life and human decisions. In this and similar contexts, it is of genuine interest to make the cause of the correlation among the answers to survey questions observable, measurable, and comparable by using the available data as a vehicle. In other cases, there is no interest in modeling what exactly caused the correlation, but quantifying the strength of the association and finding a common source of the correlation is useful for exploration and prediction. For example, people who watch the same kind of films may share a similar taste, and information on that taste can lead to better predictions of the next film they may like.

Both if the cause is of interest and if it is only a means to improve predictions, latent variable models can help. These models encode the *causes* of observed correlations in unobserved, hypothetical constructs. Latent variable models can be used to investigate the presence of underlying phenomena of either continuous (how strong is someone's religious faith?) or discrete (which religion does someone identify with?) nature.

In a nutshell, when using latent variable models, we assume that the observed cor-

relation pattern, e.g., similar answers in a survey, depends only on a few unobserved (latent) *causes*, and the "rest" of individual variance in response behavior is random and statistically independent. More precisely, the idea of latent variable models is usually formulated with a conditional distribution assumption, as explained for example by Everitt (1984): Let $\mathbf{x}^T = [x_1 \dots x_p \dots x_P]$ represent P observable random variables, for example, the answers of a person to P questions concerning religion in a survey. Then, let $\mathbf{y}^T = [y_1 \dots y_m \dots y_M]$ be the M unobserved causes of responses behavior of that same person (again, random variables). For our example, assume that $M = 2$. Let y_1 represent the strength of religious faith, and let y_2 represent the religious group of the responding person. The basic assumption of a latent variable model is that the random variables $x_1, \dots, x_P$ have a joint probability distribution conditional on the latent variables $\mathbf{y}$:

$$\phi(\mathbf{x}|\mathbf{y}).$$

So far, neither the distribution of $\mathbf{x}$ nor the distribution of $\mathbf{y}$ are specified. Suppose that the survey includes a slider bar to answer each of the questions $\mathbf{x}$, and therefore questions $\mathbf{x}$ are continuous random variables. Then, ϕ is a density function, and the unconditional density of the joint distribution is:

$$f(\mathbf{x}) = \int \phi(\mathbf{x}|\mathbf{y})h(\mathbf{y})d\mathbf{y}.$$

If the survey instead was based on a Likert scale (e.g., "Very unlikely", "Unlikely", "Likely", "Very likely") or other varieties of discrete values, ϕ would be a set of probabilities, and we would replace the integral with a product. Now the primary purpose of a latent variable model is to model the way how the observed $\mathbf{x}$ depends on $\mathbf{y}$, sometimes without knowing what $\mathbf{y}$ looks like (*exploratory* analysis), and sometimes with strong hypotheses on the form of $\mathbf{y}$ (*confirmatory* analysis). As we need to infer both ϕ and h simultaneously, we need additional assumptions on the functional form of the latent variables to reduce the parameter space. Most importantly, and the

crucial assumption in latent variable models, we assume *conditional independence*: If we know about someone's religious group and strength of religious faith, the rest of the answering behavior in the survey is random.

$$\phi(\mathbf{x}|\mathbf{y}) = \phi_1(x_1|\mathbf{y}), \ldots, \phi_p(x_p|\mathbf{y}), \ldots, \phi_P(x_P|\mathbf{y}).$$

The equation shows that given the value of the unobserved variables $\mathbf{y}$, the observed variables are stochastically independent of one another. For example, the unobserved strength of religious faith and religious group cause observed correlations in survey responses, and people who respond that they go to church more often than others also pray more often than others. Once we know $\mathbf{y}$, the question answers have a specific expected value, and anything that differs from this expected value is mostly random. Once this essential assumption is accepted, usually more assumptions reduce the search space by restricting the functional form of ϕ_p and h (Everitt, 1984).

Exploratory models scan the correlation structure for latent variables that best explain the observed correlations and provide a good fit for the independence assumption. These models can, for example, reduce observed film choice to the latent variable of consumer taste in recommender systems (compare, e.g. Koren et al., 2009). In a confirmatory analysis, the unobserved variables have a particular, parametric functional form, and additional assumptions on parameters of ϕ_p and h, and are often part of a chain of causal effects in a structural equation model (Everitt, 1984; Jöreskog, 1971).

How is the class of latent variable models related to social and geographical space? In social networks, latent variable models can be used to explore unobserved common interests among people (also called homophily) or reciprocity (people are likely to respond to one another's behavior) (e.g., Hoff, 2009). In geographical data, latent variable models and related algorithms find geophysical influences within a limited geographical area and thereby account for pollution sources without explicit information on where these pollution sources are (e.g., Pesaran and Tosetti, 2011; Demšar

et al., 2013). There is a strong duality between social and geophysical nearness. So what is state of the art in latent variable based modeling of both types of data, and what epistemic and statistical challenges and use cases arise?

This thesis contributes to the understanding of how latent variable based models are related to social and geophysical phenomena. It contributes to cross-disciplinary knowledge exchange on different endeavors that that profit from latent variable models and advances methodology with new approaches to their estimation and application. In four articles, we discuss use cases in computational social science and spatial prediction. The rest of this introductory chapter is structured as follows: Section B summarizes the four articles that treat latent variable models in social and physical space via latent variables. The articles mainly focus on the social and physical dimensions of space but have many links to other forms of statistical dependency, like temporal dependency. Together, the articles emphasize the dichotomy of geographic and social space. The first article provides some extended statistical theory on recent advances in exploratory, multidimensional latent-variable models in data mining. The second article discusses an application to social networks with methodological novelty in the domain of dynamic link prediction, supported by Prof. Dr. Sven Apel and Thomas Bock, formerly from the Faculty of Computer Science and Mathematics, Chair of Software Engineering I at the University of Passau, now Chair of Software Engineering at the University of Saarbrücken. The third article proposes an application with methodological novelty to the spatial prediction of air pollution in co-authorship with Prof. Dr. Harry Haupt and Svenia Behm, University of Passau[1]. The fourth article finally considers hypothesis testing and intercultural comparison in confirmatory factor analysis, in coauthorship with Dr. Jörg Scheffer from the Faculty of Arts and Humanities, University of Passau. The high degree of interdisciplinarity raises questions to the epistemic bases of this theses, as well as the overall contribution to the field of statistics, critical data studies, and computational social science that goes beyond the contributions of the single articles (Section C). Section D finally summarizes the thesis and provides an outlook on

[1]The third article with Svenia Behm and Prof. Dr. Harry Haupt has been published in *Spatial Statistics* (`https://doi.org/10.1016/j.spasta.2018.04.004`).

follow-up projects and future work.

B Article summaries

B.1 Space and Time in Latent Variable Models

Statistical latent variable models make one fundamental assumption: There is some hidden phenomenon like someone's taste or a trend that *causes* the correlations in the data – not the other way around. However, when the model and algorithm related to such a model starts *driving* human behavior, the logic of causation may be *inverted*. With his book on "filter bubbles", Eli Pariser popularized the question of whether recommender algorithms actively reduce the diversity of content that people are exposed to on the Internet and thereby change our culture and human mindsets (Pariser, 2012). Some descriptive approaches have shown that this reduction in diversity in contents indeed exists (e.g., Nguyen et al., 2014). These descriptive approaches cannot answer the question of whether there is an adverse effect or not – the reduced diversity might merely show that the filters are working: Everyone receives the content they like. This uncertainty needs to be reduced, considering the high potential impact of filter algorithms on our society. Therefore, mathematical explanations must complement descriptive approaches. Investigating algorithms, models, and data is a vivid field of research: *critical data studies*. Typical questions in this field are: Are there assumptions in the models and algorithms that actively reduce content diversity? Can these assumptions be eliminated, or can their negative impact be mitigated?

The first article of this thesis discusses recent progress in latent variable based modeling of phenomena in social and geographical space. We discuss latent variable models for prediction with a particular focus on the way they incorporate similarity among people, their geographic distance, and their relative positions in a social network. Due to the physical distance of people, their in-take of polluted air correlates. Due to social proximity and integration into a social subgroup, they may be exposed to the same viruses. Due to similar values and convictions, they may consume the same products. Latent variable models relate observed correlation structures in em-

pirical data to the presence of an unobserved reason or an abstract phenomenon that is not measured directly. In this case, we can retrieve physical or social distance via latent variables. We show that the models are all relational in the sense that they describe a similarity. Recent progress in latent variable modeling can handle multiple dimensions of human interaction simultaneously (e.g. Cichocki et al., 2009). He et al. (2016), for example, used both information of social embedding and geographical information to recommend appropriate locations to users. The massive amounts and the high versatility of data generated in social networks and geo-location-based services allow modeling human behavior in an unprecedented level of detail. All these developments bring new challenges for the transparency and explainability of the models. Filter bubbles arise because the output of a model and training influences the data it will be trained on in the future again. Geographic distance is different from other forms of distance as it is not subject to change due to its analysis – the difference is subtle but essential for developing models that can avoid filter bubble effects. Ignoring this difference and prioritizing predictive performance, there are some models for recommender systems that treat the "nearness" to other people in the same way. This simplification may be a reasonable choice for short-term prediction optimization, but unreasonable concerning the long-term tracking of changes in the social fabric. In the long-term, models should instead take into consideration the changes that they introduce to the correlation structure as they *encourage* people to change their behavior. Such a model would allow researchers to test mechanisms that counteract the detrimental effects of the active role of filters that could substantially enhance search and filter algorithms.

B.2 Spectral Stability in Open-Source Software Developer Networks

Open-source software (OSS) development projects often depend on a relatively small group of developers, who are accountable for a large share of code contributions and coordination efforts. These *core developers* are supported by a large number of *peripheral* contributors, who invest less time and effort individually, but together add a substantial part of the value in open-source software creation (Setia et al., 2012; Crowston and Shamshurin, 2017; Joblin et al., 2017). Together, the core

and peripheral contributors form open-source *communities*, that create significant contributions like the Linux Kernel. Within the communities, *sub-groups* or *sub-communities* arise that work on a topic or task for a limited time only (Bird, 2011). Graph mining methods can identify these sub-communities with their interaction behavior. An obvious source for data on social structures in Open-Source Software Projects is version control data. Version control software simplifies task sharing and re-integration of written code. Communication systems (e.g., issue trackers) and e-mail lists complement version control systems for coordination of the development tasks. As both data sources are publicly accessible, it is possible to track the dynamics of developer sub-communities and activity levels in both the context of communication and collaboration over several decades (Bird et al., 2006; Joblin et al., 2015).

The idea of the second article of the thesis is to develop a model to evaluate the stability of group behavior, modular structures, and social ties in open-source communities over time. For this purpose, we develop a new network-based method that involves the estimation of dynamic latent factors. Moreover, we want to track the congruence of both systems in time. We synthesize the *Additive and Multiplicative Effects* (AME) latent factor model suggested by Hoff (2009), and the *spectral growth* model suggested by Kunegis et al. (2010).

The spectral growth model extrapolates the growth of a latent dimension to future periods. It thereby explores and utilizes the dynamics in the activity level and the importance of single parts of a software development project for link prediction improvement. The *spectral growth* model (Kunegis et al., 2010) allows to assess the stability of factorial patterns. It also allows exploiting this stability to enhance prediction, when groups of developers in the network grow more important over time or tend to lose in importance.

In the AME model, the probability of one node i to have a connection with node j depends both on observed and unobserved characteristics of the node, and the interplay of those characteristics between two particular nodes. If all these pieces of information were known, then the single cooperation or communication events between two developers would be stochastically independent of the rest of the net-

work. As this is not the case, we use the observed network structure to infer the unobserved information implicitly. In the context of OSS, assume that a group of developers is very closely connected: The group's members communicate and collaborate regularly. Then these developers can be assumed to have "something" in common. A multiplicative latent factor error structure can represent this "something". If the underlying organizational structures and "reasons" to collaborate are stable over time, then the factorial structure of the network will also be stable, and only the relative importance of single patterns will change.

How can the two models be combined? The AME model can both flexibly infer different types of unobserved patterns and account for predictors and thereby measure congruence. The spectral growth model is more mechanistic and can only deal with one network as an input. However, it provides a dynamic perspective and allows for tracking the stability of unobserved patterns over time. By combining both models, we receive a tool that can both flexibly infer modular structures and exploit group dynamics in time. Additionally, the new tool explores the congruence of structural patterns in communication and collaboration networks. Beyond the assessment of stability and prediction, we use our approach to quantify the extent to which communication ties exceed *ad hoc* collaboration structures, and how stable the structure in communication is over time. For replication of the original method proposed by Kunegis et al. (2010), we use spectral decomposition for factor extraction. For the more advanced settings, we use a *Monte Carlo Markov Chain* (MCMC) based procedure.

We find, to our surprise, a relatively weak relationship between communication and collaboration, at least when considering both forms of interaction within three-month time windows. Moreover, we find weak evidence for increases in the stability of social ties over time or with a growing number of programmers that participate in a project. Against our expectations, short-term collaboration seems to be no good predictor of communication, implying that coordination efforts do not respond to coordination needs in time.

B.3 Spatial Detrending revisited: Model(l)ing Local Trend Patterns in NO_2-Concentration in Belgium and Germany

Tracking and predicting current levels of air pollution are essential to the prevention of health risks and the planning of political measures. The third article of this thesis proposes an enhanced method to make real-time predictions about local pollution, departing form air pollution data. We combine these time series with information on land use within the direct surroundings of the station, that is, information whether industrial use, housing areas, streets, or other types of use dominate the direct vicinity of a measurement station.

Two-step methods, like the one chosen in this article, seek to remove any observable trend by modeling long-term local average pollution levels and removing them. This procedure is the prerequisite for the application of spatial prediction methods (that is, *kriging* methods) that rely on the absence of trends in the first and second distribution moment (that is, the absence of heterogeneity in expected value and variance-covariance, also known as weak stationarity). We base upon a model for the removal of deterministic trends in the first and second distribution moment suggested by Hooyberghs et al. (2006). The model or method has been denominated "RIO model" for the specific context of air pollution analysis with land use data by Janssen et al. (2008). With the help of air pollution data and land-use classes, we estimate a local long-term trend. This long-term trend expresses whether the vicinity of a measurement station is dominated by industrial usage, traffic, or housing areas, with the collateral consequences for air quality. We innovate the approach by a more flexible approach to the choice of the form of this trend. For this purpose, we use an encompassing nonparametric regression, following the kernel estimation approach for mixed continuous and categorical data of Li and Racine (2004, 2007). Trends are calculated by simultaneously smoothing over a pollution vector indicator and additional explanatory covariates. Having identified a low order polynomial with the help of the nonparametric approximation, we succeed in substantially improving the cross-sectional predictive performance of the model by taking into account additional heterogeneity between urban and rural stations. Our empirical analysis reproduces the previous results of Janssen et al. (2008) for Belgium and provides evidence for

Belgium and Germany that the suggested modifications perform very well.

B.4 The Role of Blueprints in Quantitative Cultural Comparison.

The abundance of data from Social Networks, virtual communities, and digital art creation, has led to a switch from research based on "small" to "big" data in cultural analytics. The comparative sociological perspective on culture has embraced the "big data" world and the analysis of Twitter data, Instagram data, and the like (Manovich, 2018). Such data-driven approaches use exploratory quantitative approaches for the measurement of culture. This approach is not new and by no means related to the big data movement: Pierre Bourdieu with his *sociology of culture* was using exploratory quantitative analysis for inference of sociological patterns back in the 70ies already. According to his model, hidden social *milieus* determine observed behavior and habits like playing golf or consuming a specific kind of food (Bourdieu, 1977). Like in latent variable analysis, he infers a latent group (the milieu) from observed behavior.

Comparative cultural analytics based on this approach suffers from a lack of objective cultural yardsticks - because the milieu is derived from observed behavior, it can not be tested statistically with the same data. The epistemic foundation of such an analysis is very different from other theories that assume a group first, establish highly reliable cultural dimensions, and then serve for group comparison. The most prominent example for this approach is the pioneering work by Geert Hofstede (e.g. Hofstede, 2003). Instead of observing behavior and then drawing conclusions on the groups, Hofstede operationalizes the concept of *dimensions of culture*, measures these unobserved dimensions via survey questions, and compares between groups. These dimensions are related to particular persons and aggregated on the national level; for example, "the French" are compared to "the Germans". Bourdieu's approach describes groups only, not individuals, and cannot be used for comparison as it *creates* the group it is describing.

While the reduction of culture to nationality may be overly flawed, the geographical reference provides an easy to understand tool for the mediation of differences. What role do blueprint-based categorization play for comparative cultural analyt-

ics nowadays? The Bourdieu-inspired approach to derive social groups from observed behavior, and the Hofstede-inspired approach to finding objective measures and yardsticks and test hypotheses about cultural differences between pre-assigned groups. *Comparative Cultural Analytics* is a field where latent factor models have played an essential role in the past. In an exploratory way, they can identify people of similar tastes or behaviors and choices that usually come together. The confirmatory way establishes fixed measurement models with clear hypotheses on what underlying phenomena drive a particular behavior. The Hofstede-based approach requires objective yardsticks, making sure that the compared concepts have the same meaning in all contexts.

Group-based structural equation modeling is a modeling approach that verifies this presumed neutrality. The group-based model assigns a group-specific parameter structure to the relationship between observable (survey) items and the underlying psychometric concept encoded in the item correlations. The confirmatory multiple group factor model (Jöreskog, 1971) provides an inferential approach to determine a degree of invariance ("comparability") of the measurement models. Confirmatory models have to balance two competitive objectives: On the one hand, statistical approaches in the context of quantitative cross-cultural research have to be impartial and provide unbiased inferential results and the basis for hypothesis testing. On the other hand, they should not be overly restrictive, as this would impede the discovery of informative structures that go beyond clear-cut national collectives. The model-based approach with *a-priori* group assignment is still an active field in science when the aim is to establish testable hypotheses about closed groups, and the GESIS research institute collects data about human values and convictions every few years, offering them in the framework of the *European Values Study* for scientific purposes.

Article 4 focuses on the role of group-based latent factor models in cultural comparison. What role can political boundaries and well-established measurement models possibly play for the conceptualization of culture, when the exploratory mining of geotagged data becomes ever more popular? We want to fill a gap in the discussion of the relationship between clear-cut political boundaries and the empirical methodology of cultural comparison. We empirically illustrate the confirmatory approach

to modeling culture and boundaries, using data from the European Values Study (EVS). We conclude that it may be true that with globalization, traditional explanatory schemes that frame culture within fixed political (national) borders lose their adequacy for the explanation of cultural, social, and economic processes. People travel more and have the opportunity to share their thoughts and beliefs with other people around the globe. However, many elements from the Hofstede-inspired approach help to avoid understanding the risks that arise from biased data and over-interpretation of group differences. By explaining the considerations that are usually part of a "small data" approach, we seek to inspire the adoption of quality assurance measures in "big data" methodology for comparative cultural research.

C Discussion

The four articles in this thesis are joint work with colleagues from other fields of empirical research, and one of the benefits of putting them together in this thesis is to have a context-spanning discussion of their contribution. This section summarizes the epistemic context (Section C.1), as well as the overarching contribution of this thesis to the fields of computational social science (Section C.2), critical data studies (Section C.3), and latent variable modeling with spatial and network data (Section C.4).

C.1 Epistemic Context

The high degree of interdisciplinarity of this thesis comes along with an evident mixture of epistemic traditions, which started to manifest as a core question of the thesis - were the models *describing* reality, trying to infer causal relationships, or trying to optimize prediction of future events? Computational social science, as described by Conte et al. (2012), often requires the combination and reconciliation of multiple traditions, because it combines topics from social sciences with methods and models developed in statistics and computer science (Masson et al., 2017). The next paragraphs explain the epistemic context of our work concerning big data versus small data approaches, and the objective of the models in terms of descriptive,

explanatory, and predictive aims.

First of all, the four articles in this thesis differ in terms of their objectives. The articles pursue *descriptive*, *explanatory*, and *predictive* aims. As articles 1 and 4 elaborate on in more depth, latent variable models serve all three purposes. Many applications of latent variable models seek to explore the latent variable structure of observed physical and social datasets to improve *predictive* power of a model. Such algorithms contribute to tasks like forecasting, link prediction, and recommendation, by estimating an unobserved latent variable structure from observed correlation. In a similar vein, in the third article, we use land use information for enhancing spatial prediction. The link between this index and the expected pollution level is established in a non-parametric way to avoid strong assumptions on the relationship between land use classes and pollution. The choice of a specification does not depend on a causal explanation or explainability but on predictive performance only.

On the other hand, latent variable approaches are useful for customer segmentation or the detection of local sources of pollution – that is, for *descriptive* or *explanatory* aims. The explanation of a causal relationship requires an analysis to be able to generalize the interplay of a target system's components to rules on how the system behaves (Conte et al., 2012). In the second article, the article on the stability of group patterns in Open Source Software Developer social networks is mainly descriptive, even if we validate the adequacy of the model with a measurement of predictive performance. We seek to find out about the stability of the group patterns to allow project leads to compare their projects and the respective organizational structure to other projects. We validate the fit of single assumptions with a link prediction performance evaluation and cross-validation, yet we seek to interpret the decomposition of the matrices and the stability of group patterns via the decomposition. This exploratory task distinguishes the second article from the fourth article, which profits and relies on a well-designed sampling process of survey data that are designed according to well-established and validated scientific constructs like institutional religiosity. Article four is a typical *explanatory* approach: We seek to investigate and test hypotheses about cultural homogeneity and measurement model validity, ignoring how well the model would predict the cultural convictions of a single survey

participant. The explanatory implications imply that samples have to be representative - which makes them inadequate in settings like the social networks built in article 2, where the bias of the data set is unknown.

In complementation to the distinction between descriptive and predictive purposes, latent variable models are related to *inductive* and *deductive* statistical research. Inductive statistics seeks to infer properties of the population from a data sample, and the inductive statistical toolset implies interval estimates, hypothesis tests, and multivariate variance analysis. Believing in the possibility to derive population characteristics and causal rules that also hold when applied in a deductive way (that is, anticipating an observable behavior from the identified rule) means that data have to be representative in a global context (Halfpenny, 1987). Inductive data analysis often serves as input for the generation of explanatory, deductive models, and plays a crucial role (Jebb et al., 2017). Our findings on group stability in the second article could fuel deductive, explanatory research based on alternative statistical approaches, which would help to reject or support the conclusions we draw from exploratory graph analysis. Model-driven research, as discussed in article 4, profits from the discovery of new patterns in exploratory research. Both approaches can either contribute to the fair inference of other parameters, the visualization and operationalization of unobservable patterns, or the enhancement of prediction performance. An exploratory latent variable analysis depends only on very few prior assumptions, and it reduces multiple covariates to fewer aggregate patterns that make the correlations interpretable. In this way, exploratory latent variable analysis is helpful when the actual driving mechanisms behind preferences, social ties, and other latent features are unknown, or when their specification is too costly. This thesis contains examples of both inductive and deductive objectives. For the sake of completeness, it should be mentioned that a third paradigm profits from both inductive and deductive research, and is a core paradigm in Computational Social Science: generative approaches. To gain insights into social systems when data is noisy or not available, researchers use simulations to anticipate the behavior of a hypothetical system. The agent-based modeling uses simulation approaches and projects specific assumptions on behavior and properties into data analysis. The

generative approach constitutes a valuable alternative to the descriptive component of article 2 and 4, as it has been used both in cultural analytics (e.g. Axelrod, 1997; Flache and Macy, 2011) and social network analytics (e.g. Block et al., 2019). Agent-based modeling is nowadays widely used, for example, in network modeling to test assumptions about how people interact (e.g., Block et al., 2019). The agent-based modeling uses simulation approaches and projects specific assumptions on behavior and properties into data analysis (Conte et al., 2012).

Third, the articles differ in their interpretation and importance of the assumption that the latent factors are *causal* to the observed correlation patterns. Latent variable models assume that we can infer the causes of correlation patterns from observed data. In a Humean sense, causality is "constant conjunction", that is, the two phenomena are consistently observed together and can be described with covariation patterns, and contributes to inductive reasoning with inferential statistics (that is, statistics based on probability theory and hypothesis testing) (Halfpenny, 1987). The discussion of causality is essential for the interpretation of the outcomes of a latent variable analysis - often, researchers rely on their causal intuitions to give a name to a latent variable, like "religiosity". In confirmatory latent variable models like the ones discussed in article 4, the causality hypothesis, as well as strong assumptions on the structure of this causality, are essential for hypothesis testing. For example, only if it really *is* social solidarity and nothing else that causes differences in response behavior, the measurement model is valid and suitable for hypothesis testing. While some models *assume* causality, they are unable to *proof* it – experimental research tries to do that. Other real-world phenomena may have caused the correlations than those that we assume, two phenomena that overlay each other, and other sources of spurious correlation may occur. In article 3, we do not care whether a land-use class *causes* air pollution - we just measure whether the correlation between land use class and pollution is strong enough to improve spatial prediction.

C.2 Contributions to Computational Social Science

Articles 1, 2, and 4 are rooted in the field of Computational Social Science, the research on social phenomena with computational methods. The discussion on the epistemic foundations of quantitative research in social sciences has gained impetus with the progress in *Big Data* analysis in the last few decades. In the *Manifesto of computational social science*, Conte et al. (2012) explain the many possibilities that arise from the usage of data generated by information and communication technology (ICT), to analyze social phenomena that were not feasible before. The authors argue that ICT produces a "flood of data"(Conte et al., 2012, p.327) that leaves us with "traces of almost all kinds of activities of individuals enabling an entirely new scientific approach for social analysis" (Conte et al., 2012, p.327). Together with the progress in computational efficiency and capacity, this allows researches to use an unprecedented variety of models and an unprecedented depth of detail in social diversity and complexity.

Scaling computational methods like latent variable modeling from a small data world to a big data world exposes data analysis to the "new epistemologies of data science"(Kitchin and P. Lauriault, 2015, p.464). For the investigation of social phenomena, many risks arise if spurious correlations are over-interpreted, if researches neglect bias in data generation, or if the output of the model has an effect on the training data and therefore has a self-reinforcing automatism.

> Tackling a question through big data often means repurposing data that were not designed to reveal insights into a particular phenomenon, with all the attendant issues of such a maneuver, for example creating ecological fallacies (Kitchin and P. Lauriault, 2015, p.466).

Masson et al. (2017) observe that the adoption of Big Data tools in research has pushed back interpretive and critical perspectives, and "humanistic scholarship seems to get increasingly indebted to positivist traditions"(Masson et al., 2017, p.25). Underlying phenomena can be measured and analyzed via the measurement. Article 4 and the idea of being able to infer cultural patterns from observed behavior, as suggested by Bourdieu early had intensive debates with adherents of constructivism,

who argue that it is impossible to collect data on culture without influencing the result as the designing researcher is subject to his own cultural bias.

The sampling process for Big Data like social network data we mine in article 2 usually involves no controlled sampling processes, and therefore Big Data contain biases, noise, and abnormality. A core challenge that this thesis reveals is how to distinguish between *small data* and *big data* application cases of latent variable models and what this means for the generation of knowledge. Kitchin and P. Lauriault (2015) discuss the role of small data models in a world of ever-better data availability, data pooling, sharing, and linking. Structural equation models like the one discussed in article 4 stem from a small data world, where data was produced in a controlled way. Sophisticated stratigies served to fight bias, uncertainty, and errors in data (Miller, 2010). The aim to verify that this endeavor has been successful is one of the core topics of article 4, with hypothesis tests. These strategies do not apply for large datasets generated by Information and Communication Technology, as used in article 2. Neither can it be entirely guaranteed for the application case of article 3, it is hard to validate whether measurement stations have been put at specific places to assess, for example, predominantly the pollution exhibition of people in city centers. Big Data are usually a "by-product of systems rather than being designed to investigate particular phenomena or processes"(Kitchin and P. Lauriault, 2015, p.463).

When neglecting the specific needs of big data research, empirical research is prone to four types of "fallacies" (Kitchin, 2014, pp.133–137). First, some researchers treat big data as if they could provide a holistic, gap-free picture of reality in full resolution, that is, to an arbitrary degree of detail. It is unreasonable to assume that big data can provide a more complete, higher resolved picture of religiosity of people, as often, there are gaps in the data, and the data generating process is not well designed. In a similar vein, the analysis of article 2 is valid only to the extent that the version control data reflect the actual programming process, but provide no full picture. Second, big data analysis often pretends that there is no need for a priori theory, models, or hypotheses. Exploratory latent variable models and descriptive methods often depart from a descriptive approach, avoiding the disadvantages of inferential

statistics. That makes it harder to use them to validate a theory, as shown in article 2. In article 3, the most flexible setting, the one with the non-parametric function definition, is the one that produces relative much instability and is in the way of an interpretation of the pollution index. Third, big data analytics fails if it assumes that data "can speak for themselves free of human bias or framing"(Kitchin, 2014, p.136). Bias is the topic of article 4 – while the confirmatory we use there seeks to prevent bias via inferential tests, many newer approaches ignore data bias. In article 1, we just live with the bias in the data and try to describe the data that is there as precisely as possible. Fourth, Kitchin (2014) asks the same questions that researchers in culture have been asking for decades – it is a misconception according to some epistemic traditions that meaning and interpretation can be transferred from one domain of knowledge to another. When various epistemic traditions are mixed, the scientific consequences, benefits, and disadvantage, merit thorough evaluation.

C.3 Contributions to Critical Data Studies

The field of *Critical Data Studies* comprises methodological work on how to avoid the fallacies of data-driven research. Kitchin (2017) provides a structured introduction into critical data studies, distinguishing between "thinking critically about" algorithms and "researching" algorithms (Kitchin, 2017, p.16). The author underlines the need for empirical investigations of algorithms and the long-term consequences of their application, and emphasize the highly contextual way algorithms perform. Barocas et al. (2013) and Kitchin (2017) suggest six different perspectives for critical data studies. Three of these perspectives apply to our research: First, algorithms can be studied from a *technical* (computer science) perspective, explaining their implementation, algorithmic steps, optimization steps, and interaction with hardware components. We follow the technical perspective in articles 1, 2, and 3, with the detailed technical description of the methods we are implementing (we emphasize, however, more the statistical steps than the technical steps). Second, we use a *philosophical* approach by raising the question of algorithm ethics in article 1, as well as some epistemic questions in article 4. Third, we use the *socio-technical* perspective to discuss how algorithms shape particular domains of research.

In article 2, we seek to infer rules of developer behavior from observed interaction patterns, in an exploratory way. Accepting that our data only cover small parts of actual coordination behavior, we measure the fit of specific assumptions on human behavior to the data and derive insights into the stability of social bonds. In article 3, we do not seek for rules, but merely for enhanced predictive performance. The air pollution data in the third article follow a particular design with non-random sampling (e.g., pollution is measured in city centers more frequently). We approached the generation of a local pollution index from a non-parametric perspective, not imposing a parametric frame on the relationship between the general pollution level and the land-use class index. In article 4, finally, we base on small data that were specifically designed and sampled to overcome the suspicion of being subjective, to find general measurement models that are comparable across groups, and bring objectivity into a measurement.

One of the challenges for *statistical research* in this context is to find ways to make bias visible and algorithms transparent, to avoid adverse long-term effects on society. As Schäfer and Van Es (2017) put it, in a "datafied society", the main challenge of the field of data-supported research in humanities regarding algorithms is to "develop an understanding of the mathematical concepts and models driving these programmes not in order to fully master them but rather understand them sufficiently enough to approach new research objects from a critical perspective" (Schäfer and Van Es, 2017, p.16). Only by understanding the underlying assumptions that a statistical method makes on human behavior, it is possible to adapt its functionality to commonly accepted social norms. Ramsay (2003) explains the difference between seeing statistical methods as a means of supporting research in humanities for being more objective, as advertised, for example, by Hockey (2000), and understanding the role of the researcher in feeding assumptions, hypotheses, and opinions into a statistical analysis.

As Kitchin (2017) summarizes, there are three main approaches to the critical discussion of algorithms: First, a detailed case study of a single algorithm, or class of algorithms; second, a detailed examination of the use of algorithms in one domain; or, third, a more general discussion of the nature and performance of algorithms.

By raising the question of how the model assumptions are related to recent concerns on the effect of filtering (often based on latent variable models) affects society, article 1 raises an ethical discussion on the epistemic basics of big data analysis and thereby contributes to the field of *critical data studies*. From a sociological perspective, the algorithms have changed the way we consume information and interact with one another, replacing the act of searching by passive recommendation. Efficient algorithms for the factorization of sparse matrices and tensors are a prerequisite to lifting the value of massive data that online interaction continually produces: They filter the information we receive and largely determine our consumption behavior and social contacts. Hence, a proper understanding of these algorithms is fundamental in appraising the cultural development of our society.

Whether embracing big-data analytics is beneficial to social sciences or not is controversial. Some researchers in humanities fear that "[b]y succumbing to the lure of scientism [...] humanists run the risk of forgetting what they excel at – critical interpretation – and by the same token, of impoverishing their practice"(Masson et al., 2017, p.26). Over the years, however, data-driven research in the humanities evolved from a supporting tool to a research area of its own, which lives from trans-disciplinary cooperation of different fields of research (Masson et al., 2017). The field of *Critical Data Studies* departs from the assumption that *big data* and associated algorithmic processes shape the society as well as the individual. *Big data* is in this context understood as a phenomenon driven by digitization and describes the inflow of uncontrolled data. *Small data* describes a controlled way of data, also involving considerations on the data generating process (Kitchin and P. Lauriault, 2015).

The four articles in this thesis, with their four individual epistemic bases, reflect different perceptions about the subjective nature of data, and how they are related to human perception and behavior. Manovich (2018) traces how the mathematical description of human behavior has undergone multiple stages of paradigms. Before the mid-nineteenth century, deterministic models of human behavior were prominent in sociology. Similar to physical laws, human behavior was believed to be determined by mathematical rules. The probabilistic method of describing behavior

became the dominant way of thinking about human behavior only by the end of the nineteenth century–when social scientists started investigating correlations in measurable behavior and started adopting concepts of dependent and independent variables (Manovich, 2018). Confirmatory latent variable models, as described in article 4, describe the aim to describe fixed rules of human behavior in probabilistic terms: the most crucial assumption of the model is that the answer to a question in the structural equation model depends on an unobserved factor. The data that the validation of this model depends on is highly controlled and curated, and the model contains strong assumptions on the statistical properties of the data. These assumptions are invalid in the big-data context, where data and sampling bias are basically always there. Manovich (2018) suggests that big data is a new paradigm in social sciences: Instead of relying on a limited set of variables for many people, we now have an unlimited number of variables to analyze a particular person. There is no need to restrict the number of categories in comparative analysis because Big Data provide endless possibilities in describing social behavior in more dimensions than the usual time and geographic dimension plus a few research-specific dimensions like literature. This kind of analysis can enrich our understanding of culture (Manovich, 2018).

At the same time, it is essential to keep the limitations of such big data studies in mind. Following the categorization of approaches to *critical data studies* suggested by Kitchin (2017), we cover several perspectives on algorithms and data: First, we explain core assumptions on big data algorithms in article 1 (category 5), covering the politics and power embedded in algorithms. We suggest that the filtering used for big relational datasets within several contexts explores correlations, and the application of the algorithm leads to more homogeneity in recommendations. Cutting into the ethical perspective of algorithms (category 1), we criticize the measurement of the "success" of recommender systems, which is inseparable from the application of the algorithm. It remains unclear whether a good fit of the clicks means that the algorithm fits reality well, or that the algorithm changes people's behavior. With article 2, we contribute to the basis of perspective 2, the sociological approach. Our approach extends the toolset for looking into the creating of software and the role

of stable social groups in this approach. Article 3 follows perspective number 1, improving the performance of an algorithm regarding its predictive power. Indeed, the article shows the compromise between ad hoc assumptions such as a linear relationship between land use classes and air pollution and increased flexibility through non-parametric specification. The computational cost competes with performance, measured by a simple output-related performance measure. We thus followed the recommendation of Kitchin (2017) to combine more than one of the approaches.

C.4 Contributions to Statistics and Latent Variable Modeling

From a statistical, methodological perspective, this thesis contributes to the field of latent variable modeling with social and geographical data. Articles 2, 3, and 4 are all applied research and shed light on different aspects of latent variable modeling. In article 2, we develop a new way to extrapolate factorial structures in social networks. Social networks consist of people that interact – nodes and edges. When using statistical models on network ties, we need to account for human behavior that creates dependency among the network ties. For example, homophily means that people that are "similar" are more likely to interact. This similarity can be a shared hobby, a shared home town, or equal gender. It is not always possible to observe the source of similarity that leads to this behavior, or at least it becomes very tedious to model every mechanism separately. Latent variable models can be used in this context to prevent problems with breaches of independence assumptions and to improve prediction. By applying a network decomposition with a dynamic extrapolation of weights, we provide an innovative way to deal with dependency in network data pragmatically, in between the descriptive and predictive paradigms.
In article 3, we propose a non-parametric specification search for mapping spatial covariates to predict a a spatial output. While we do not explicitly work with factors here, our prediction utilizes a land-use indicator that linearly combines land-use class-specific pollution to a single pollution indicator. It thereby reduces land use classes to an unobserved local pollution level (similar to latent variables). It also demonstrates the difference between geographic space and network space. All positions, as well as distances, are assumed to be exogenous and do not change. The

innovation of our approach consists in the combination of the specification search and spatial extrapolation, which improves spatial air pollution prediction.
All models discussed above are exploratory, using very few prior assumptions on the latent variable's structures. Alternatively, in some contexts, there are strong theories on how to measure an unobserved phenomenon, such as a cultural value. In this context, assumptions on the correlation structure and its homogeneity are essential when reliable estimates are needed. Article 4, joint work with the field of anthropological geography, we question the practice of fixing cultural in- and out-groups for cultural comparison. We show how explanatory, deductive research with strong assumptions on the statistical properties of a model relates to the practice of modern quantitative cultural research, and what kinds of problems can arise in this context.

D Conclusion and Outlook

This thesis provides four different perspectives on the use of latent variable models, which encode the presence of some unobservable phenomenon. The overarching contribution of the articles comprises the fields of Computational Social Sciences, Critical Data Studies, as well as methodological research in Latent Variable Modeling.
In follow-up studies with colleagues from the field of Empirical Software Engineering, we seek to apply multivariate network analysis to a similar kind of question: How do technical and social systems influence each other? And how can the insights on this interaction of technical and social systems used to optimize collaboration practice among developers? In the context of agile software development, the interplay of communication and collaboration is (compared to traditional working environments) of increased importance (compare Dybå and Dingsøyr, 2008). In Empirical Software Engineering, the "mirroring hypothesis" describes the phenomenon there is a close interconnection between an organization's structure and the structure of the resulting software products, such that the resulting products "mirror" the organization's communication and collaboration structures (compare MacCormack

et al., 2012). The definition of communication structure can impede more informal communication ways. The technical structure constrains the choice of developers – just like the pre-selection of filter algorithms constrain the choice of users. By combining qualitative with quantitative perspectives on the software development process, we model the interplay of different roles in a developer team, with different modes of interaction. Especially multi-modal network modeling will be useful for predicting and understanding developer behavior and group dynamics in real-world organizations.

Such studies of complex social phenomena can profit from interdisciplinary exchange, as potential drawbacks in methods manifest in interdisciplinary discussions mainly. Questioning model assumptions, potential bias in parameter inference, and the epistemic basis of such analysis can lead to more valuable insights into social processes.

E References

Axelrod, R., 1997. The dissemination of culture: A model with local convergence and global polarization. Journal of Conflict Resolution 41, 203–226. doi:10.1177/0022002797041002001.

Barocas, S., Hood, S., Ziewitz, M., 2013. Governing Algorithms: A Provocation Piece. Technical Report. Available at SSRN: https://ssrn.com/abstract=2245322. doi:10.2139/ssrn.2245322.

Bird, C., 2011. Sociotechnical coordination and collaboration in open source software, in: 2011 27th IEEE International Conference on Software Maintenance (ICSM), 568–573. doi:10.1109/ICSM.2011.6080832.

Bird, C., Gourley, A., Devanbu, P., Gertz, M., Swaminathan, A., 2006. Mining email social networks, in: Proceedings of the 2006 International Workshop on Mining Software Repositories, ACM, Shanghai, China. 137–143. doi:10.1145/1137983.1138016.

Block, P., Stadtfeld, C., Snijders, T.A., 2019. Forms of dependence: Comparing

saoms and ergms from basic principles. Sociological Methods & Research 48, 202–239.

Bourdieu, P., 1977. Outline of a Theory of Practice. Volume 16. Cambridge university press.

Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.i., 2009. Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons, Chichester, UK.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M., Helbing, D., 2012. Manifesto of computational social science. The European Physical Journal Special Topics 214, 325–346. doi:10.1140/epjst/e2012-01697-8.

Crowston, K., Shamshurin, I., 2017. Core-periphery communication and the success of free/libre open source software projects. J. Internet Services and Applications 8, 10:1–10:11. doi:10.1186/s13174-017-0061-4.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S., McLoone, S., 2013. Principal component analysis on spatial data: an overview. Annals of the Association of American Geographers 103, 106–128. doi:10.1080/00045608.2012.689236.

Dybå, T., Dingsøyr, T., 2008. Empirical studies of agile software development: A systematic review. Information and Software Technology 50, 833–859. doi:10.1016/j.infsof.2008.01.006.

Everitt, B.S., 1984. An INtroduction to Latent Variable Models. Springer Science & Business Media. doi:10.1007/978-94-009-5564-6.

Flache, A., Macy, M., 2011. Local convergence and global diversity: From interpersonal to social influence. Journal of Conflict Resolution 55, 970–995. doi:10.1177/0022002711414371.

Halfpenny, P., 1987. Laws, causality and statistics: positivism, interpretivism and realism. Sociological Theory 5, 33–36.

He, J., Li, X., Liao, L., Song, D., Cheung, W.K., 2016. Inferring a personalized next point-of-interest recommendation model with latent behavior patterns, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 137–143.

Hockey, S.M., 2000. Electronic texts in the humanities: principles and practice. Oxford University Press.

Hoff, P.D., 2009. Multiplicative latent factor models for description and prediction of social networks. Computational and Mathematical Organization Theory 15, 261–272. doi:10.1007/s10588-008-9040-4.

Hofstede, G., 2003. Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. Sage publications, Thousand Oaks, USA.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., 2006. Spatial interpolation of ambient ozone concentrations from sparse monitoring points in Belgium. Journal of Environmental Monitoring 8, 1129–1135. doi:10.1039/b612607n.

Janssen, S., Dumont, G., Fierens, F., Mensink, C., 2008. Spatial interpolation of air pollution measurements using CORINE land cover data. Atmospheric Environment 42, 4884–4903. doi:10.1016/j.atmosenv.2008.02.043.

Jebb, A.T., Parrigon, S., Woo, S.E., 2017. Exploratory data analysis as a foundation of inductive research. Human Resource Management Review 27, 265 – 276. doi:https://doi.org/10.1016/j.hrmr.2016.08.003.

Joblin, M., Apel, S., Hunsen, C., Mauerer, W., 2017. Classifying developers into core and peripheral: An empirical study on count and network metrics, in: Proceedings of the 39th International Conference on Software Engineering, 164–174.

Joblin, M., Mauerer, W., Apel, S., Siegmund, J., Riehle, D., 2015. From developer networks to verified communities: a fine-grained approach, in: Proceedings of the 37th International Conference on Software Engineering (ICSE), IEEE Press. 563–573.

Jöreskog, K., 1971. Simultaneous factor analysis in several populations. Psychometrika 36, 409–426. doi:10.1007/BF02291366.

Kitchin, R., 2014. The data revolution: Big data, open data, data infrastructures and their consequences. Sage.

Kitchin, R., 2017. Thinking critically about and researching algorithms. Information, Communication & Society 20, 14–29. doi:10.1080/1369118X.2016.1154087.

Kitchin, R., P. Lauriault, T., 2015. Small data in the era of big data. GeoJournal 80, 463–475. doi:10.1007/s10708-014-9601-7.

Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. Computer 42, 30–37. doi:10.1109/MC.2009.263.

Kunegis, J., Fay, D., Bauckhage, C., 2010. Network growth and the spectral evolution model, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), 739–748. doi:10.1145/1871437.1871533.

Li, Q., Racine, J., 2004. Cross-validated local linear nonparametric regression. Statistica Sinica 14, 485–512. doi:10.2307/24307205.

Li, Q., Racine, J., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press, Princeton, USA.

MacCormack, A., Baldwin, C., Rusnak, J., 2012. Exploring the duality between product and organizational architectures: A test of the "mirroring" hypothesis. Research Policy 41, 1309 – 1324. doi:10.1016/j.respol.2012.04.011.

Manovich, L., 2018. The science of culture? social computing, digital humanities and cultural analytics. URL: osf.io/preprints/socarxiv/b2y79, doi:10.22148/16.004.

Masson, E., Schäfer, T., van Es, K., 2017. Humanistic Data Research: An Encounter between Academic Traditions. in: Schäfer and Van Es (2017). 25–37.

Miller, H., 2010. The data avalanche is here. shouldn't we be digging? Journal of Regional Science 50, 181–201. doi:10.1111/j.1467-9787.2009.00641.x.

Nguyen, T.T., Hui, P.M., Harper, F.M., Terveen, L., Konstan, J.A., 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity, in: Proceedings of the 23rd International Conference on World Wide Web (WWW), ACM. 677–686. doi:10.1145/2566486.2568012.

Pariser, E., 2012. The filter bubble: How the new personalized web is changing what we read and how we think. Penguin Books, London, UK.

Pesaran, M.H., Tosetti, E., 2011. Large panels with common factors and spatial correlation. Journal of Econometrics 161, 182–202. doi:10.1016/j.jeconom.2010.12.003.

Ramsay, S., 2003. Toward an algorithmic criticism: Reconceiving text analysis. Literary and linguistic computing 18, 167–174.

Schäfer, M.T., Van Es, K., 2017. The Datafied Society: Studying Culture through Data. Amsterdam University Press.

Setia, P., Rajagopalan, B., Sambamurthy, V., Calantone, R., 2012. How peripheral developers contribute to open-source software development. Information Systems Research 23, 144–163.

Technical Implementation in R

All statistical computation in this thesis has been carried out in the open-source software R (R Core Team, 2013), and I would like to thank all voluntary maintainers of the packages that I used. Some important general R packages for my work are ggplot2 (Wickham, 2009), knitr (Xie, 2017), rmarkdown (Allaire et al., 2017) and xtable (Dahl, 2016). More details can be found in articles/chapters 2,3, and 4. Central R packages for article 2 are igraph (Csardi and Nepusz, 2006) for network objects, and eigenmodel (Hoff, 2012) and amen (Hoff et al., 2015) for MCMC-based estimation of the *Multiplicative and Additive Effects* model, and RSpectra (Qiu et al., 2016) and irlba (Baglama and Reichel, 2015) for the decomposition of large matrices. The analysis is embedded in work of Prof. Dr. Sven Apel (University of Saarbrücken) and his team, published as R-package CoRoNet. The R packages we use for the analysis in article 3 comprise broom (Robinson, 2017), GISTools (Brunsdon and Chen, 2014), gstat (Pebesma, 2004; Gräler et al., 2016), np (Hayfield and Racine, 2008), optimx (Nash and Varadhan, 2011; Nash, 2014), raster (Hijmans, 2016), rgdal (Bivand et al., 2017), spatstat (Baddeley et al., 2015), and timeDate (Rmetrics Core Team et al., 2015). The most important packages used for article 4 are lavaan (Rosseel, 2012), psych (Revelle, 2014), semPlot (Epskamp, 2014), semTools (semTools Contributors, 2016) and sp (Bivand et al., 2013).

F References

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., Hyndman, R., Arslan, R., 2017. rmarkdown: Dynamic Documents for R. URL: https://CRAN.R-project.org/package=rmarkdown. R package version 1.10.

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press, London, UK.

Baglama, J., Reichel, L., 2015. irlba: Fast Truncated SVD, PCA and Symmetric Eigendecomposition for Large Dense and Sparse Matrices. URL: https://CRAN.R-project.org/package=irlba. R package version 2.0.0.

Bivand, R., Keitt, T., Rowlingson, B., 2017. rgdal: Bindings for the Geospatial Data Abstraction Library. URL: https://CRAN.R-project.org/package=rgdal. R package version 1.2-7.

Bivand, R., Pebesma, E., Gómez-Rubio, V., 2013. Applied Spatial Data Analysis with R. Use R!, Springer, New York, USA.

Brunsdon, C., Chen, H., 2014. GISTools: some further GIS capabilities for R. URL: https://CRAN.R-project.org/package=GISTools. R package version 0.7-4.

Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. Complex Systems 1695, 1–9.

Dahl, D.B., 2016. xtable: Export Tables to LaTeX or HTML. URL: https://CRAN.R-project.org/package=xtable. R package version 1.8-2.

Epskamp, S., 2014. semPlot: Path diagrams and visual analysis of various SEM packages' output. URL: http://CRAN.R-project.org/package=semPlot. R package version 1.0.1.

Gräler, B., Pebesma, E., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. The R Journal 8, 204–218.

Hayfield, T., Racine, J.S., 2008. Nonparametric econometrics: The np package. Journal of Statistical Software 27, 1–32. doi:10.18637/jss.v027.i05.

Hijmans, R.J., 2016. raster: geographic data analysis and modeling. URL: https://CRAN.R-project.org/package=raster. R package version 2.5-8.

Hoff, P.D., 2012. eigenmodel: Semiparametric factor and regression models for symmetric relational data. URL: https://CRAN.R-project.org/package=eigenmodel. R package version 1.01.

Hoff, P.D., Fosdick, B., Volfovsky, A., He, Y., 2015. amen: Additive and Multiplicative Effects Models for Networks and Relational Data. URL: https://CRAN.R-project.org/package=amen. R package version 1.1.

Nash, J.C., 2014. On best practice optimization methods in R. Journal of Statistical Software 60, 1–14. doi:10.18637/jss.v060.i02.

Nash, J.C., Varadhan, R., 2011. Unifying optimization algorithms to aid software system users: optimx for R. Journal of Statistical Software 43, 1–14. doi:10.18637/jss.v043.i09.

Pebesma, E.J., 2004. Multivariable geostatistics in S: The gstat package. Computers and Geosciences 30, 683–691. doi:10.1016/j.cageo.2004.03.012.

Qiu, Y., Mei, J., Guennebaud, G., Niesen, J., 2016. RSpectra: Solvers for Large Scale Eigenvalue and SVD Problems. URL: https://CRAN.R-project.org/package=RSpectra. R package version 0.12-0.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, AT. URL: http://www.R-project.org/.

Revelle, W., 2014. psych: Procedures for Psychological, Psychometric, and Personality Research. URL: https://CRAN.R-project.org/package=psych. R package version 1.4.8.

Rmetrics Core Team, Wuertz, D., Setz, T., Chalabi, Y., Maechler, M., Byers, J.W., 2015. timeDate: Rmetrics – chronological and calendar objects. URL: https://CRAN.R-project.org/package=timeDate. R package version 3012.100.

Robinson, D., 2017. broom: convert statistical analysis objects into tidy data frames. URL: https://CRAN.R-project.org/package=broom. R package version 0.4.2.

Rosseel, Y., 2012. lavaan: An R package for structural equation modeling. Journal of Statistical Software 48, 1–36. doi:10.18637/jss.v048.i02.

semTools Contributors, 2016. semTools: Useful tools for structural equation modeling. URL: https://CRAN.R-project.org/package=semTools. R package version 0.4-14.

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer, New York. URL: http://ggplot2.org.

Xie, Y., 2017. knitr: A General-Purpose Package for Dynamic Report Generation in R. URL: https://CRAN.R-project.org/package=knitr. R package version 1.15.1.

Chapter 1

Space and Time in Latent Variable Models

Chapter Abstract

With an automated judgment on preferences, filter algorithms provide people with an individual choice of content and contacts. Thereby, they facilitate navigation within virtual social communities, search engines, and even physical surroundings. This service is, however, suspect to reduce the diversity of contents and information that people have access to, leading to adverse social effects. While this effect has been described empirically by checking the diversity of contents, a description does not provide insights on the underlying reasons for identified changes.

One way to get these insights is to mathematically explore the related models: Latent variable models as a specific class of recommender algorithms that should be investigated more closely in this regard. Technological progress allows latent variable algorithms to take ever more dimensions into account simultaneously. We show that many recommender methods that involve latent variable models are explicitly specified in a way that pretends to merely *describe* unobserved patterns. However, there are related inferential models that provide a good starting point for a quantitative assessment of the strength of the influence of recommender algorithms. The article thereby opens new perspectives for the integration of ethical aspects in information filtering.

1.1 Introduction

Recent progress in computational capacity has opened many new possibilities for processing multi-dimensional data. Multi-dimensionality means that data involve multiple different types of attributes, like a time dimension, a geophysical position, and individual characteristics like age. Another dimension on this list is a social, relational dimension. The processing of masses of relational data that result from the mass adoption of social networks as a core means of exchange adds a fundamentally different aspect to data than individual attributes than can be retrieved in traditional market research surveys. Together with the widely adopted usage of GPS trackers and recommendation algorithms for shopping, multi-dimensional data describe a person in an unprecedented depth of detail.

Algorithms that process these data are designed to optimize the predictions they make – who is someone likely to befriend, where is someone likely to go, and what is someone likely going to buy. The algorithms process information on what, whom, and which places someone likes without *influencing* what some likes – preferences are supposed to be *exogenous* or *independent* in a model or method. Many recommendation and prediction algorithms either assume this exogeneity of data or do not involve any considerations about it. However, what if this assumption is wrong, and the algorithms *influence* our preferences instead of just detecting them? The "filter bubble" hypothesis, popularized in 2012 by Eli Pariser, says that people can get "trapped" in the pre-selected content that the Internet offers to them. This trap can have negative consequences for the cohesion of society because filter algorithms have a self-reinforcing effect. While human brains have been filtering information ever since they existed, Pariser argued that the situation is different now because the creators of the algorithms actively *want* people to get trapped in filter bubbles (Pariser, 2012). Nguyen et al. (2014) conducted a longitudinal study on the effect of collaborative filtering on the diversity of contents that people had access to in their virtual environment. The authors found that, indeed people's information consumption is influenced by such algorithms. Recent research confirms that exposure to ideologically versatile content on social media can be restricted if no

counter-measures are in place (compare Bakshy et al., 2015).

There is an ethical imperative to investigate the relationship between statistical models and social phenomena more closely. Pariser diagnosed the transition from human gatekeepers like the editors of newspapers, geographic scope, and physical boundaries of knowledge exchange, to digital, algorithmic gatekeepers to content and information. Therefore, the author argues, algorithms need to be provided with a "a sense of civic responsibility" (Pariser, 2012). This imperative becomes even more pressing when considering that "big data" related methods and models are increasingly also entering the field of *descriptive* and *explanatory* research in computational social sciences (Manovich, 2018). There are many advantages of doing so: big data provide endless possibilities in describing social behavior in more dimensions than the usual time and geographic dimension plus a few research-specific dimensions like literature. This kind of analysis can enrich our understanding of culture (Manovich, 2018).

Time, geographical position, and social network are the three dimensions that most directly affect people in their preferences. Physical proximity has lost in importance for social interaction in a digitized society: "glocalization" describes the phenomenon that nowadays, individuals engage in multiple interest-based virtual communities that have both a local embedding and a global extent (Johnston and Pattie, 2011). Onnela et al. (2011) investigated the interplay of physical proximity and connectivity in social networks and social exchange. With the help of a community detection algorithm, the authors clustered people in virtual communication networks and found the effect of geographic proximity on the intensity of social exchange to decline with peer group size. In other words: The most intimate groups of people are still mainly local, while the radius of exchange grows with the number of nearest persons one considers. Strong social ties are closely related to local proximity.

This article discusses the underlying assumptions on people's behavior of some recently developed models and methods for prediction and recommendation. We put particular emphasis on three dominant dimensions of interpreting human behavior: time and geographical position, as well as social network. With this article, we contribute to the interdisciplinary exchange between social science, predictive

Time: Changes in the other dimensions, temporal influences like time of the day	**Social Network:** Represents cluster structures, topology of networks, group dynamics, relations	**Feature space:** Describes unobservable mental characteristics, beliefs and attitudes	**Geographical space:** Describes the embedding of nodes, ties and bonds in geographical space

Figure 1.1: Four dimensions of human preferences.

analytics, and statistical inference. We discuss the representation of social and geophysical change in recommender systems. For our mathematical consideration, we chose a class of models that enjoys widespread acceptance in both geographical as well as network-based analysis, and which contains an active modeling perspective and differentiation between spatial modes – latent variable models. For computational reasons, factorial models used to be applied to two dimensions at a time only. However, recent methodological and technical progress made it possible to explore geographical embedding and social embedding simultaneously, through the use of higher-order factor models for the analysis of data from *Location Based Social Networks* (LBSNs).

Our results suggest that while some algorithms actively "claim" to be descriptive only (by implying certain independence conditions), they cannot assess their effect on the training data basis due to the way they are specified. There are, however, inferential approaches in the field of higher-order factorial models that can infer change. More research is needed to come up with models that can do both at the same time. The remainder of the article is organized as follows. Section 1.2 defines fundamental statistical concepts in the context of *latent factor models*, like *space*, *mode*, and *dimension*. We outline a basic two-dimensional factor model and explain the way such models treat different notions of space. In 1.3, the discussion of two-dimensional (i.e. matrix-based) data structures is extended to higher-order (i.e. tensor-based) data structures. Finally section 1.5 summarizes the findings of the review and concludes.

1.2 Latent Variables in Two-Dimensional Spatial Models

In section 1.2, we introduce the class of latent variable based models and factorization-based recommender algorithms. This class of model/method is widely spread for recommendation ever since the Netflix competition for recommending films to platform users (compare Koren et al., 2009). We start with an explanation of the statistical background of the models and emphasize the difference between the treatment of geographical space and social space.

1.2.1 Factor Models and Factorization

Suppose a researcher is confronted with a survey dataset, including a hundred person's answers to the questions "How often do you go to church?", "How often do you pray to God?" and "How often do you consult the Bible?". Intuition would say that those questions reflect how *religious* a person is. It is possible to relate the three questions to a joint underlying concept – *religiosity*, the unobserved "driver" of the answers. This reduction of three observed variables to a single joint variable is also called *dimensionality reduction*: Instead of considering all three items simultaneously, it is possible to reduce them to a single score on an artificial "religiosity"-scale. If "How often" can be answered on a continuous scale, the latent construct is usually assumed to be normally distributed.

The 100 people's answers to the three questions mentioned above will likely correlate: People that often go to church may also be more likely to consult the Bible often. Both questions are dependent, and their answers are not independent of one another. However, two equally religious people may have the same expected response to both questions - any deviations from those expected values are, however, random. In statistical terms, this means that the answers to both questions are *conditionally independent* - once the underlying factor is known (that is, if the actual religiousness of two people is observable), the answers are identically and independently distributed. The strong correlations are the result of the presence of an underlying pattern or construct, and therefore, vice versa, the correlations can be used to infer those unobserved constructs.

To formalize this insight, suppose there is a set of observable random variables (i.e. questions asked or *items*) x_p ($p \in 1, \ldots, P$) with means μ_p. Moreover, let F_k ($k \in 1, \ldots, K$) denote a set of unobserved random variables or *factors*. These factors represent the underlying lower-level concepts that determine the responses, like the underlying reasons for the response behavior of a survey participant. There are fewer factors than original variables, i.e., $K < P$. Item-specific factor loadings l_{pk} are used to describe the importance of factor F_k for item x_p. The model bases on the assumption that the unobserved underlying factors reveal themselves partially in the observed variables. The larger the absolute value of l_{pk}, the stronger the relationship between a factor and a variable. Formally, we get

$$x_p - \mu_p = l_{p1}F_1 + \ldots + l_{pK}F_K + \epsilon_p, \tag{1.1}$$

where ϵ_p denotes an independently distributed error term with zero mean and finite variance-covariance structure. The distribution of ϵ_p may vary across p. Figure 1.2 visualizes this idea: common underlying factors cause the observed correlation structures. A cross-sectional set of the responses of N survey participants to a survey involving P items, the answers to which are driven by K unobserved factors (usually $K < P$) can be expressed in matrix notation:

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{F}\mathbf{L}^T + \mathbf{E}, \tag{1.2}$$

where $\mathbf{X}$, μ and $\mathbf{E}$ are of order $N \times P$, $\mathbf{F}$ of order $N \times K$, and $\mathbf{L}$ of order $P \times K$. Each column of $\mathbf{F}$ contains one unobservable factor $\mathbf{f}_k$ (e.g. the N persons' *religiosity*, *honesty*, or whatever unobserved concept may influence the person's answering behavior), and a single row in $\mathbf{F}$ contains person n's factor values within the K unobserved dimensions. $\mathbf{L}$ is called the matrix of factor loadings. Any model that fulfills the structure above is a latent factor model. A general introduction to latent factor models can be found, for example, in the work of Loehlin (2004).

Given the factor values, the observable data is independently and identically dis-

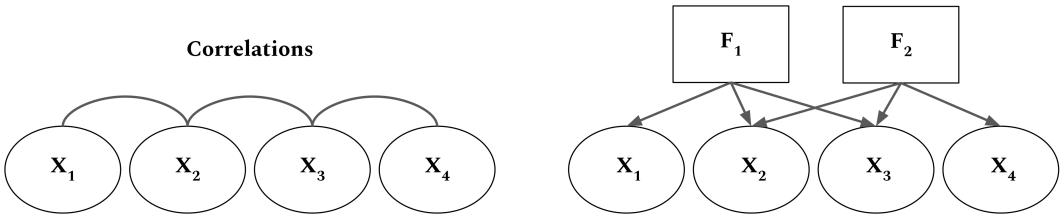

Figure 1.2: Representation of the basic latent factor model: observed correlations results from the presence of unobserved factors.

tributed:

$$\mathbf{X}|\mathbf{F} \sim i.i.d. \tag{1.3}$$

It depends on whether $\mathbf{L}$ is assumed to be deterministic or random whether it is necessary to condition on $\mathbf{L}$. As Jolliffe (1986) explained, it is usual to assume that the factors are random variables, and the loadings are fixed. The factors (the columns of $\mathbf{F}$) are restricted to be orthonormal (mutually orthogonal and normalized to a length of 1), formally $\mathbf{F}\mathbf{F}^T = \mathbf{I}_N$. Consequently, the variance-covariance matrix of the observable items $\mathbf{X}$ can be expressed as

$$VCov(\mathbf{X}) = E((\mathbf{X} - E(\mathbf{X}))^2) = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi} \tag{1.4}$$

The two components that add up to the variance-covariance matrix, $\mathbf{L}\mathbf{L}^T$ and $\mathbf{\Psi}$, are called *communalities* and *specificities*, respectively (Loehlin, 2004; Jolliffe, 1986). Generally, the formation of factors from highly correlated items via linear combination, without prior hypotheses on the structure of such factors, is referred to as *exploratory factor analysis*. There are different ways to specify the model (for example, the orthogonality restriction can be suspended). In our discussion, we focus on the interplay of different dimensions in the exploratory analysis of correlation structures. Survey data describes *features* and *individuals*.

The correlation matrix can either reflect attributes that show high correlation (several items represent the concept of religiosity) or reflect clusters of people that show similar response behavior (a group of people tends to give the same answers). On the one hand, it is possible to use a survey among individuals to calculate the correlations of single items and determine which items can be combined linearly to a lower-dimensional information sub-space. The factors are then linear combinations of items, which express underlying "reasons" for correlation between the items. On the other hand, $\mathbf{X}$ can be transposed to $\mathbf{X}^T$, to calculate the correlations between people to determine how similar answers are among people. Latent factors are then linear combinations of individuals, forming groups. Items of features with strong correlations can be clustered into highly correlated *groups of features*, as they express the same unobservable *factors*, and thereby the factors become measurable. In order to group people, their similarity or distance in feature space has to be estimated. Distance can be measured by Euclidean distance in an unlimited number of attributes (Demšar et al., 2013).

As shown above, factorization applies to correlation structures and finds patterns without a proper model. The results of a factor model estimation or a factorization identifies similar individuals. Similar individuals are in this case either regions with similar pollution patterns or people with similar religiosity. Factorization is usually model-free and applied "mechanically", whereas the factor model requires assumptions on distributions and properties of the data, and allows for inference of parameters. Before starting to explain the differences between the settings, we would like to underline the value of using a model and having inferential ambitions, instead of mechanically applying factorization. One of the risks in the factor-based analysis of spatial structures is the overinterpretation of point estimates. Dommenget and Latif (2002) discussed the risk of mistaking principal components and orthogonal factors for *meaningful* patterns. The authors named several examples from highly renowned journals that identify physically important patterns in climatology from the second principal component. However, as the authors showed, the interplay of several *patterns* or *modes* in climate lead to a superposition of all patterns in the first principal component, and an artificial antagonism in the second principal component. Similar

problems have been described by von Storch (1999a), who concentrated on "misuse" of supposedly sophisticated methods of analysis for finding spectacular results. The model forces factorial structures to be orthogonal, while there is no reason to believe that real-world processes follow this rule. Rather than interpretation, PCA concentrates on the efficient representation of variance. In order to discriminate between spurious and true patterns, Zwiers and von Storch (2004) suggested to trace the identified patterns in time and check their dynamic behavior for plausibility and reproducibility. A possible remedy against overconfidence in estimates is taking account of the differences in stochastic and deterministic factors formulation, and respecting the fact that several factorial structures can lead to the same observed patterns. For a probabilistic formulation of models with accurate factor inference, Bayesian methods are the tool of choice. Probabilistic frameworks include possibilities of penalized estimation in sparse data environments (see Salakhutdinov and Mnih, 2007), and Bayesian analysis in combination with a Monte Carlo Markov Chain (MCMC) finally allows for the inclusion of prior information through the choice of an adequate prior distribution (Hogan and Tchernis, 2004; Salakhutdinov and Mnih, 2008; Zhu et al., 2016). The probabilistic formulation of blockmodels for network analysis (Airoldi et al., 2008) and for the unobserved latent social space (Hoff et al., 2002; Handcock et al., 2007) can help to avoid erroneous conclusions from clustering in the context of latent factor analysis in a network context.

1.2.2 Factorization of Geographical Space for Pattern Recognition

Matrices reflect two dimensions at a time, one along the rows, and one in columns. When the answers of N persons to one single question are tracked over T periods, the matrix is of dimension $N \times T$ and reflects the two modes: individuals and time. Time, or "*dynamic* space" is one possible *mode* of factor analysis. Time-related factors describe common trends that affect all individuals in different intensities, where intensity manifests in factor loadings. Geographical space is a fourth possible mode of analysis. Referring to the three subspaces geographic space, temporal space, and attribute space, Richman (1986) and Jolliffe (1987) defined six modes of factor analysis via *Principal Component Analysis* (PCA), which are summarized, for

example, in Demšar et al. (2013)[1]. The model seeks to measure either an underlying pattern in one or the other dimension, depending on which dimension relates to the factors and which to the loadings. In the *O-mode* the relationship between *features* (attributes/items/characteristics) and *time* is modeled. In this mode of factor analysis, phenomena developing in time, like an economic downturn, can be modeled and used for prediction. When the order is inverted, the *P-mode* expresses the need to cluster features, whose development in time is similar. The *Q-mode* clusters similar locations regarding attributes, and the *R-mode* clusters attributes that are similar at the same locations. The *S-mode* identifies latent patterns in locations where events take place at the same time, and the *T-mode* identifies patterns in time where events happen at the same locations. Table 1.2.2 provides an overview of the purpose of the single modes in direct juxtaposition.

Matrix factorization is not only an exploratory tool for the discovery of patterns in observations that violate the independence assumption. It is also used to estimate the error structure in other models, which would usually depend on an independent error structure. As Pesaran and Tosetti (2011) pointed out, the *multifactor* approach is, besides the explicit inclusion of neighborhood structures as suggested for example in Wang and Wall (2003), one of the two dominant approaches for dealing with error cross-section dependence in panel data where N is large relative to T (that is, when there are more individuals than time stamps). While the usage of neighborhood matrices requires hypotheses on which individuals interact with one another, the factor-based approach is exploratory, and "discovers" spatial dependencies by itself. The multifactor approach assumes that a finite number of unobserved common factors can describe the cross-sectional dependency structure in the data. Anything beyond the observed effects (that is, the error term) can be decomposed into an idiosyncratic error and the combination of a few common effects (Pesaran and Tosetti, 2011). For spatiotemporal data, assume y_{it} is an observed outcome for individual i at time t. Then the interplay of dynamic factors and individual factor loadings determines y_{it}. As the distribution of the factor loadings is unrestricted,

[1]PCA is one basic method for factor extraction. In contrast, factor models generally use assumptions that go beyond the dimensionality reduction in PCA.

Table 1.1: Modes of multivariate analysis are defined by the combination of feature space, geographic space, and time (e.g., Demšar et al., 2013). The horizontal line delineates an extension of the overview by a social network perspective.

Mode	Loadings (**l**) vs. Factors (**F**)	Description
O-mode	Attributes vs. Time	Attributes are considered data elements and sampling times variables; changes in features are due to phenomena developing in time, such as trends and seasons.
P-mode	Time vs. Attributes	Sampling times are data elements and attributes are variables; similar changes in time are due to common underlying characteristics of individuals.
Q-mode	Attributes vs. Locations	Attributes are considered data elements, and locations are variables; measurements are similar because a set of locations has similar characteristics.
R-mode	Locations vs. Attributes	Locations are data elements and attributes are variables; similar attributes are clustered. Also called *spatial objects PCA* or *raster PCA*.
S-mode	Time vs. Locations	Sampling times are considered data elements and locations are variables; nearby phenomena develop similarly over time. Also called *atmospheric science PCA*.
T-mode	Locations vs. Time	Locations are data elements and sampling times are variables; time effects drive changes in the location of objects (concerns phenomena that can change location).

the factors loadings can show spatial clusters, for example, when nearby regions are similar. While the common factors model temporal variation, the factor loadings take account of the spatial variation and dependency (Abadie et al., 2010; Lopes et al., 2011; Pesaran and Tosetti, 2011). To stay in the context of human interaction, consider the following example: The political climate in a country can change over time, where y_{it} describes the dominance of pro-European voices. Every country has a $(1 \times K)$ vector of factor loadings, which describes the reaction to K different temporal influences, which can be general developments or time shocks that affect all countries equally. In every single period t, K such factors influence y_{it}. These

influences are encoded in the time-dependent factors. The European sovereign debt crisis and its side-effects, for example, represent a dynamic factor that affected all European countries. However, the pro-European opinion changed differently due to various degree of economic impact in the single countries and different cultural backgrounds. For example, when economic resilience is distributed unequally from North to South across Europe, then the output y_{it} is also going to show differences in the impact of an economic crisis from North to South. The strength of the impact is in this case encoded in the factor loadings.

Spatial factors describe regionally limited phenomena that cause spatial heterogeneity in the output. Therefore, a core purpose of spatial factor models is to decompose the analyzed area into regions that represent a spatial signal, that is, a phenomenon that is characterized by a spatial covariance structure that acts predominantly over a particular spatial scale, indicating a spatial factor (Bailey and Krzanowski, 2000). Geographical factor analysis seeks to identify regions where similarity is high, and thereby intends to find informative signals in the data. The unobserved drivers for local heterogeneity reveal themselves in variance-covariance structures that are homogeneous over a limited area of a network or space, that is, the area which is affected by the spatial signal. Correlations that go beyond a certain radius are associated with noise and considered to be spurious. The separation of signals from noise can be helpful, for example, to clean satellite imagery data (Bailey and Krzanowski, 2000).

One of the first formal factor-based models in the geospatial context has been proposed by Switzer and Green (1984). The model utilizes a continuous spatial index $\mathbf{s}$. Switching to a continuous index is useful because spatial data is not ordered in space, and describing a circular region around a spot s of radius δ is easier in continuous indexing than in discrete indexes. The following explanations are based on the work by Bailey and Krzanowski (2000). The model emphasizes that it is not the dyadic relationship between two individuals that creates dependencies, but their geographic proximity. Let

$$\mathbf{y}(\mathbf{s}) = (y_1(\mathbf{s}), \ldots, y_p(\mathbf{s}))^T,\ \mathbf{s} \in \mathcal{R} \tag{1.5}$$

be a collection of p spatial processes, with second-order stationarity, i.e. the p processes emanate from a random process with constant mean, and spatial covariance depends only on distance and direction separating any two locations. $\mathbf{s}$ is a continuous set of point locations in spatial domain or *region* $\mathcal{R}$. According to Switzer and Green (1984), the process is driven by K latent factors $u_k(\mathbf{s}),\ k = 1, \ldots, K$. The authors defined the process structure in terms of covariance and correlation:

$$u_k(\mathbf{s}) = \mathbf{a}_k^T \mathbf{y}(\mathbf{s}),\ k = 1, \ldots, K,\ \text{ such that} \tag{1.6}$$

$$Cov(u_k(\mathbf{s}), u_j(\mathbf{s})) = Cov(u_k(\mathbf{s}), u_j(\mathbf{s} + \boldsymbol{\delta})) = 0,\ k \neq j\ \text{ and} \tag{1.7}$$

$$Cor(u_k(\mathbf{s}), u_k(\mathbf{s} + \boldsymbol{\delta})) \leq Cor(u_j(\mathbf{s}), u_j(\mathbf{s} + \boldsymbol{\delta})),\ k < j. \tag{1.8}$$

$\boldsymbol{\delta}$ is some suitably chosen predetermined, fixed spatial separation, or lag. Next let $\boldsymbol{\Sigma} = var(\mathbf{y}(\mathbf{s}))$ be the overall dispersion matrix of these processes. The spatial factor coefficients $\mathbf{a}_k$ are the normalized eigenvectors of $\boldsymbol{\Sigma}^{-1}\mathbf{C}(\boldsymbol{\delta})$, where $\mathbf{C}(\boldsymbol{\delta}) = Cov(\mathbf{y}(\mathbf{s}), \mathbf{y}(\mathbf{s} + \boldsymbol{\delta}))$. The eigenvectors, representing spatial factors, are mutually orthogonal and therefore uncorrelated. The more eigenvector/eigenvalue pairs one assigns to the factorial structure, the more autocorrelation within the area defined by the lag $\boldsymbol{\delta}$ is covered by the factor structure. The patterns that represent a smaller share of the total variance-(auto-)covariance structure, i.e. the factors with the smallest eigenvalues, can be assumed to represent noise only. The idea behind this argument is that $\mathbf{y}(\mathbf{s})$ can be composed of $\mathbf{y}^{(1)}(\mathbf{s})$ and $\mathbf{y}^{(2)}(\mathbf{s})$, which might represent signal and noise. Therefore, the Variance-Covariance-Structure can also be decomposed into a signal and a noise component, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)}$, where $\boldsymbol{\Sigma}^{(k)} = var(\mathbf{y}^{(k)}(\mathbf{s}))$ (Switzer and Green, 1984; Bailey and Krzanowski, 2000).

Orthogonal spatial factors resulting from spectral decomposition are commonly referred to as *Empirical Orthogonal Functions* (EOFs). In a discrete formulation, these

represent the spatial equivalent of a Principal Component Analysis, while it can also be combined with a spatially continuous formulation via the spatial indicator **s** as shown above (Cressie and Wikle, 2011). Additional material on EOF analysis and its relationship to principal components can be found in Jolliffe (1986), von Storch (1999b), and Cressie and Wikle (2011).

A core field of research relying on EOFs and spatial factor models is the field of *Factorial Ecology* (Demšar et al., 2013). In an early definition of the field, Berry (1971) defines three main characteristics of factorial ecology: the *use of factorial methods* in an *ecologic context* with *comparative aspirations*. PCA can be one method of factor extraction. Demšar et al. (2013) showed that nonspatial PCA, which is bases on the assumption of independent data, is in regular use in different fields of spatial data analysis. The authors provide an overview of use cases for different kinds of spatial data and explain nonspatial, kernel-based techniques to correct for spatial analysis in PCA. The *geographically weighted PCA* technique assumes that there are regions of geographical space with distinct variance-covariance structures, which should be modeled separately. Observations are weighted inversely proportional to geographical (usually Euclidean) distances between a particular point and its neighboring data points. The resulting PCA structures vary continuously over the entire geographic space.

A less mechanical, more model-based approach to spatial factor analysis has been provided first by Christensen and Amemiya (2001, 2002, 2003). The authors developed a semiparametric latent variable model for rectangular grids, formalize assumptions of the factorial model, and provide guidance for inference. Model-based approaches to spatial factors were the basis for the development of Bayesian methods that specify parametric distributions for the single parameters (e.g., Wang and Wall, 2003), hierarchical spatial factor models (e.g., Hogan and Tchernis, 2004) and Bayesian spatiotemporal models (e.g., Lopes et al., 2011).

1.2.3 Factorization of Social Space for Link Prediction

Factor analysis in social networks seeks to identify a subspace (that is, a region or cluster) where people share a common variance-(auto-)covariance structure due to

some unobserved signal. In the same way, we can factorize social space to identify structures like groups and communities. Virtual space in terms of a network-based space of human interaction describes relationships – how close two persons are to one another can be determined, for example, by the number of people a message would have to cross to get to the other person. The current distance between two persons is determined by the path between two nodes in a network at time t.

In the spatial analysis of network structures, the spatial structure is considered to be known, as in the case of a pre-defined, invariant neighborhood matrix. Instead, we consider the spatial distance between two nodes to be the path length between two nodes and use it for link prediction. Consider a network of senders and receivers of messages, as suggested by Hoff (2009). Then the respective graph G exhibits a directed *adjacency matrix* $\mathbf{Y}(G)$. This implies that it is possible that $y_{ij} \neq y_{ji}$. y_{ij} indicates whether there is a connection between the nodes i and j. If a network describes people that talk to one another, $\mathbf{Y}(G)$ is of dimension $N \times N$, and $y_{it} = 1$ if one person i talks to another person j, and $y_{ij} = 0$ if not. In the *unipartite* setting, all nodes can be connected, and the adjacency matrix is squared. In a *bipartite* setting,there are two separate sets of nodes and no connections between nodes belonging to the same set. Consequently, the adjacency matrix is rectangular. For example, a movie platform connects N users with F films, where $y_{ij} = 1$ if person i has seen the film j. In bipartite networks, there are two separate sets of nodes, which can not be connected to another member of the same group. Figure 1.3 illustrates the difference.

Nodes are also called vertices, and ties are called edges or links (and, when they have a direction, arcs). It might seem to be an obvious choice to model the binary question whether there is a relation or not via generalized linear models like Logit models. However, such models assume nodes to be independent given the observable predictors of tie formation. There are several obvious examples that point to a regular violation of this assumption. For example, the propensity of two vertices to connect depends both on the activity level of one as well as on the activity level of the other node: Two very active nodes are more likely to connect than two nodes that hardly have any connections at all. Moreover, common interests

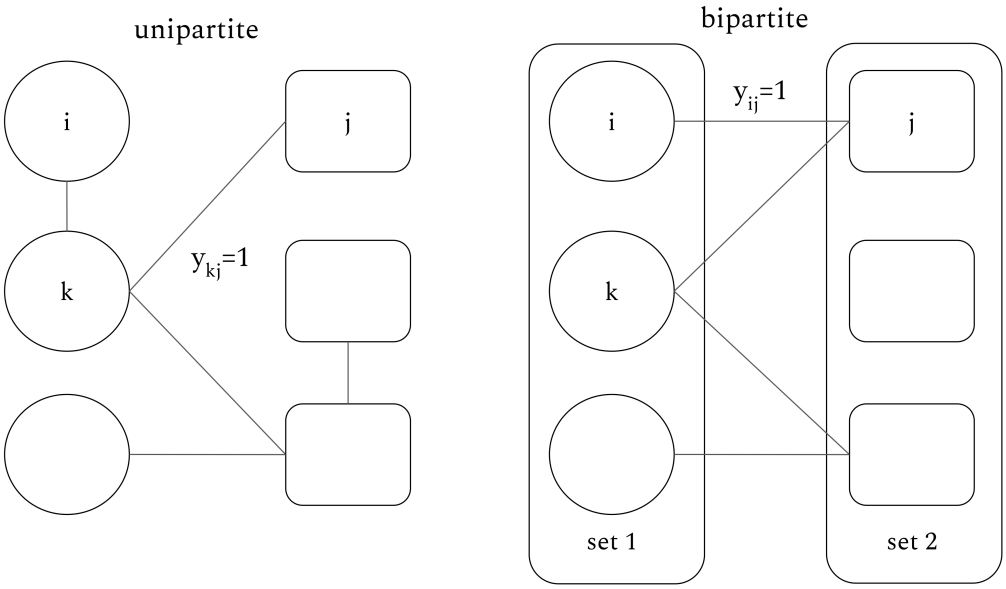

Figure 1.3: Graphical representation of unipartite and bipartite network structures.

and the exposure to common influences foster violations of independence. Only conditioning on *both* the observed and unobserved characteristics of the two nodes leads to conditional independence. Consider the probability of observing a specific network, $P(y_{ij}, y_{ik}, \ldots, y_{lk}|\theta_{ij}, \theta_{ij}, \ldots, \theta_{lk})$, where the binary variable y_{ij} indicates the presence ($y_{ij} = 1$) or absence ($y_{ij} = 0$) of a connection between nodes i and j. θ_{ij} contains all the relevant observable and unobservable information on this connection. Thus, the single edges y_{ij} are conditionally independent of one another, given θ_{ij} (Hoff, 2009; Minhas et al., 2016a):

$$P(y_{ij}, y_{ik}, \ldots, y_{lk}|\theta_{ij}, \theta_{ik}, \ldots, \theta_{lk}) = \quad P(y_{ij}|\theta_{ij}) \times P(y_{ik}|\theta_{ik}) \times \ldots \times P(y_{lk}|\theta_{lk}) \quad (1.9)$$

$$P(\mathbf{Y}|\theta) = \prod_{\alpha=1}^{n\times(n-1)} P(y_\alpha|\theta_\alpha) \quad (1.10)$$

The underlying assumption is that a small number K (i.e., $K << N$) of factors can explain the relationship between any two nodes. If the nodes describe users and films, then the factors correspond to film characteristics that determine if a user

likes the films. Usually, nobody seeks to interpret the factors explicitly. Nevertheless, they represent "obvious dimensions such as comedy versus drama, amount of action, or orientation to children [or] less well-defined dimensions such as depth of character development or quirkiness" (Koren et al., 2009, p.43) or completely uninterpretable patterns. Every user has a vector of factor values $\mathbf{u}_i$ that determines how well she suits a movie j with latent characteristics $\mathbf{v}_j$ (Koren et al., 2009). a possible representation of θ_{ij} contains observable dyadic and monadic information $\mathbf{x}_{ij}$ (such as gender if available), unobserved monadic information concerning only one of the nodes independently of the other, a_i and a_j, and a set of unobservable dyadic characteristics that interact with the characteristics of the "partner" node with corresponding weights that measure importance, $\mathbf{u}_i^T \mathbf{D} \mathbf{v}_j$. Formally, this involves

$$\theta_{ij} = \beta^T \mathbf{x}_{ij} + a_i + a_j + \mathbf{u}_i^T \mathbf{D} \mathbf{v}_j + \epsilon_{ij} \tag{1.11}$$

where $\mathbf{x}_{ij}$ denotes exogenous covariates, a_i and a_j describe random effects and $\mathbf{u}_i$ and $\mathbf{u}_j$ are vectors of node specific characteristics. Both represent *rows* in the matrices $\mathbf{U}$ and $\mathbf{V}$, i.e. $\mathbf{u}_i = \mathbf{U}_{(i\cdot)}$ and $\mathbf{v}_j = \mathbf{V}_{(j\cdot)}$. The diagonal matrix $\mathbf{D}$ contains the weights. Let z_{ij} be the factorial structure of the equation:

$$z_{ij} = \mathbf{u}_i^T \mathbf{D} \mathbf{v}_j + \epsilon_{ij}, \tag{1.12}$$

or, in matrix notation,

$$\mathbf{Z} = \mathbf{M} + \mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{V}^T + \mathbf{E}. \tag{1.13}$$

Figure 1.4 shows an example. There is a network with one observable characteristic x, which indicates two types of nodes A and B (represented by circles and rectangles). Node m is of type B, node n is of type A, i.e. $x_m = B$ and $x_n = A$. The information

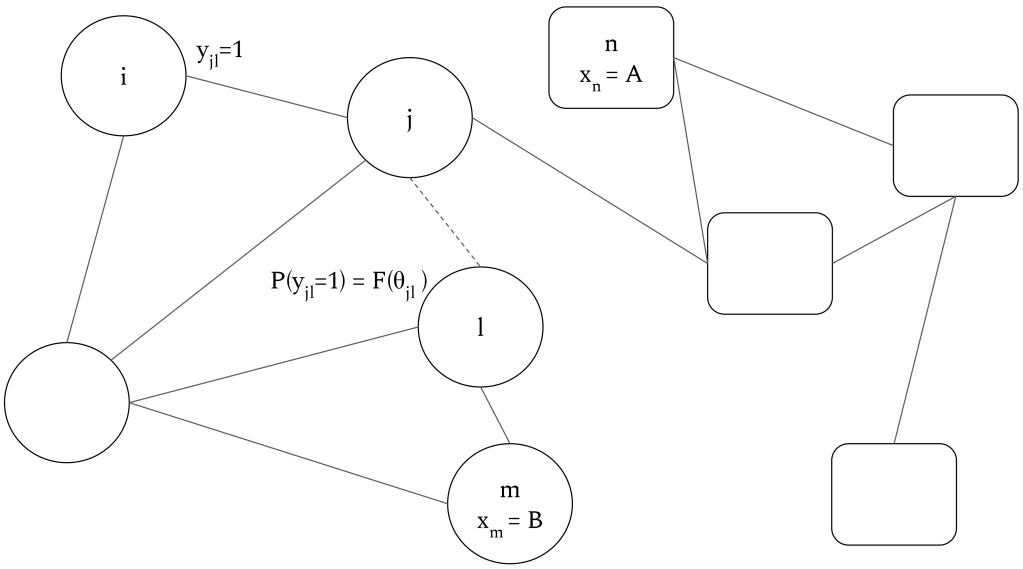

Figure 1.4: General idea of network analysis with a single binary observable variable **x**. Circles represent one category, rectangles another category.

that the nodes show stronger or more connections within a single node type is valuable information for the estimation of $P(y_{jl} = 1)$. Other clustering effects result from unobserved heterogeneity, which the latent factorial structure accounts for as well. Different extraction methods such as spectral decomposition, maximum likelihood, or Monte Carlo Markov Chain (MCMC) based methods generate matrices $\mathbf{U}$, $\mathbf{V}$ such that their columns, i.e., the factors, are orthonormal. The diagonal matrix $\mathbf{D}$ determines the importance of a factor in the tie formation process of the network (Koren et al., 2009; Hoff, 2009).

As Sidiropoulos et al. (2017) showed, discipline-specific labeling of the model expresses the various application contexts of the model. The authors distinguished research in *machine learning* from research in *signal processing*. In the latter, researchers typically focus on the columns of $\mathbf{U}$ and $\mathbf{V}$, i.e., the associated rank-1 factors $\mathbf{U}_{(\cdot i)}\mathbf{V}_{(\cdot j)}$, and the outer product of the two assigns a score for the compatibility of the two factors to every pair. This score is represented by θ_{ij}. In signal processing, research focuses on the separation of dynamic signals, which correspond to a column as they are related to time. In the former case, i.e. machine learning, researchers focus on the rows of $\mathbf{U}$ and $\mathbf{V}$, i.e. $\mathbf{U}_{(i\cdot)}$ and $\mathbf{V}_{(j\cdot)}$. These vectors are a

parsimonious representation of the position of an individual observation in a latent lower-dimensional space, like coordinates (Sidiropoulos et al., 2017). In machine learning, the interpretation of the factors is seldom of interest. Most approaches identify common factors without actually labeling them. Filter methods that explore all ratings of all users to predict a single user's rating are called *collaborative filtering* approaches (Koren et al., 2009).

Collaborative filtering with matrix factorization has become computationally very efficient around the years 2007-2010 due to the mining of sparse data structures and usage of penalty terms to reduce overfitting (e.g., Koren et al., 2009). It can now deal with very large problems that other algorithms cannot deal with anymore in a reasonable time. Matrix factorization based prediction algorithms combine "good scalability with predictive accuracy" (Koren et al., 2009, p.44). The high efficiency is also due to the split of computation and prediction in two phases - a computation and a prediction phase, which makes the prediction step independent of later updates. For fast updates of recommendations, the dimensionality reduction and clustering in the lower-dimensional subspace can be conducted in an *offline* and an *online* phase (Goldberg et al., 2001). In the offline phase, the time-consuming inference of factor structures is conducted. The "online" phase of projecting new participants into the clusters reduces to looking up values in a table, thanks to the separation of estimation and prediction. While more information on the new user is being collected, there is no need to update the entire model. Ratings adapt to the ratings of users that have rated similar films before. Instead, updates of the factorization are conducted in regular time intervals, adapting to computational capacity and size of the problem. There is no need to have all the data on a single node included in the estimation process (Goldberg et al., 2001). In contrast to other recommendation algorithms, matrix decomposition is also highly flexible about the inclusion of additional, observable information (Koren et al., 2009).

1.3 Extensions to Higher-Order Factorial Models

All the concepts discussed so far relate to two-dimensional data structures. During the last ten years, higher-dimensional approaches, based on tensor data structures, have come to maturity (compare Cichocki et al., 2009). These higher-dimensional approaches allow considering several spatial dimensions simultaneously, allowing the application of similar concepts like matrix factorization and factor-based modeling in the context of multi-modal, higher-order recommendation. For higher-dimensional analysis, data has to be organized in tensors, i.e. *multi-way arrays* or *multi-dimensional matrices*. For example, if there are five yearly survey waves with the same respondents, then the survey data can be organized in a $100 \times 3 \times 5 = N \times P \times T$ tensor. x_{npt} is then the answer of person n to question p at time t. Dimensions are also known as *ways* or *modes* of a tensor. If $I_1, I_2, \ldots, I_N \in \mathbb{N}$ denote index upper bounds, i.e. the maximum numbers a variable can take, the tensor of order N is represented by $\underline{\mathbf{Y}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$. *Scalars*, *vectors* and *matrices* are *zero-order*, *first-order* and *second-order* tensors, respectively, and tensors of order three and higher are called *higher-order* tensors (Cichocki et al., 2009).
In the next section, we discuss multi-dimensional factorial models and factorization. We start with latent blockmodels, which are directly related to the model explained in Eqs. (1.12) and (1.13). Subsequently, we discuss higher-order factorial models and then discuss models that directly refer to LBSNs. The latter type of model has a special role because it directly links social structures to the assessment of geographical positions via GPS coordinates.

1.3.1 Mixed Membership Latent Blockmodels

The models described above infer K unobserved patterns in geographic or network data. In the context of social interaction, a natural extension for latent variable modeling is to ask about group membership. Moreover, group models are among the first fields that applied tensor factorization. Hence, we start the multi-dimensional section of this article with blockmodels. As pointed out in the introduction, many people are nowadays part of several communities at a time. Latent blockmodels are

different from latent factor models as they assume the latent variables to be nominal – an individual can either be part of a community or not. It is the membership in one block of locations or in one block of nodes that determines the variance-covariance structure of an individual node or observation. The hypothesis that individuals are nowadays simultaneously member in multiple inter-driven virtual communities has a direct equivalent in *Mixed Membership Blockmodels* (MMB). Moreover, spatial clustering often uses the assumption that a location or a node is part of a group, and group membership is the only dependency structure that individuals engage in. For these reasons, we start the multi-dimensional considerations with Mixed Membership Blockmodels (MMB).

In general, blockmodels that there are unobserved groups among nodes, where the members of a group share a joint dependency structure. If the groups were known, it would be possible to include the dependencies in explicit form. For every group, one explicit variance-covariance structure applies. It is possible to use group structures in geographical applications of principal components (Demšar et al., 2013), or to include them explicitly in mathematical descriptions of a network (e.g., Wang et al., 2013). However, group memberships can also be unobserved. Latent blockmodels are similar to latent variable models and seek to infer the group structure in an exploratory way from data. *Deterministic* latent group models assume the groups to be exogenous to the system. *Stochastic* latent group blockmodels (Wang and Wong, 1987; Snijders and Nowicki, 1997) assume these groups to be stochastic, i.e., they assign a probability distribution to the membership of nodes in a group. Nodes within the class are assumed to have a homogeneous probability of connecting, and a homogeneous dependency structure and variance (Kolaczyk and Csárdi, 2014; Ferligoj et al., 2011). An overview of deterministic and stochastic blockmodels can be found in the contribution of Ferligoj et al. (2011).

The latent blockmodel makes some fundamental assumptions about space. Every node is allowed to belong to one cluster at a time only, and nodes within a cluster are stochastically equivalent, i.e., two nodes have the same probability distribution of their ties to other units (Ferligoj et al., 2011). A core difference between physical networks and virtual networks is that humans usually interact within several par-

allel network spaces at a time, by contributing to several social network platforms (Johnston and Pattie, 2011). Blockmodels can be directly related to the latent variable models discussed above when individuals are allowed to belong to several groups simultaneously. While stochastic blockmodels assign each individual to exactly one group, the Mixed Membership stochastic Blockmodel (MMB) suspends this requirement (Airoldi et al., 2008). Every individual is subject to a set of "latent roles which govern the objects' relationships with each other" (Airoldi et al., 2008, p.1982). In the MMB, each individual is assigned a vector of posterior probabilities $\boldsymbol{\pi}_i = (\pi_{1i}, \dots \pi_{Ki})$ to belong to any of the K available groups. The estimation of these probabilities utilizes a Bayesian model that involves iterative drawings from a Dirichlet distribution for the mixed membership vector. The MMB assumes that, within a group, nodes or individuals are exchangeable. It can thus be interpreted as an application of the so-called latent Dirichlet allocation proposed by Blei et al. (2003). The latent Dirichlet allocation is a generative approach for the modeling of latent classes and other discrete relationships. Initially, it has been used for complexity reduction in text analysis, in particular for the modeling of "bag-of-words", representing the hypothesis that the order of words in a document can be neglected. Airoldi et al. (2008) emphasized the close relationship between latent space models Hoff et al. (2002); Handcock et al. (2007) and the MMB. Within the MMB, it is assumed that community memberships are drawn from a Dirichlet distribution, whereas the distribution governing the interaction weights can be chosen freely. In latent space models, both latent vectors and interaction weights are drawn from Gaussian distributions.

In 2011 still, Ferligoj et al. (2011) listed the extension of latent blockmodels to dynamic models among the "open general problems for blockmodeling" (Ferligoj et al., 2011, p.443). The authors emphasized that the structural description of a network is useful as long as the structure is also indicative over time, i.e., if it remains the same over the observation period. To understand structural changes, it is necessary to include them in the model:

> If a social structure as a network really is changing then it is the fundamental structure that is changing, with the observed changes being

> indicators of the underlying fundamental change. [...] Mere descriptions of the changes involved, even if couched in terms of blockmodels, seem insufficient. We need to understand the processes generating structural change and this implies understanding how blockmodels, as representations of positions and role systems, evolve (Ferligoj et al., 2011, p.445).

A "mere description of the changes involved" alludes to the possibility to analyze single time-slices separately, as used to be state of the art until the development of dynamic models. With autoregressive elements, it became possible to relate one time-slice to the other, but until ten years ago, no integrated model existed (Kunegis et al., 2010). Fu et al. (2009) propose a dynamic state-space approach as a wrapper for the MMB proposed by Airoldi et al. (2008): the *Dynamic Mixed Membership Blockmodel* (DMMB). In the context of evolving networks, the approach serves to "dissect the evolving functional composition of the actors", using data on the interaction of an email network. Fu et al. (2009) suggested to "superimpose a state space model on top of the [MMB], and connect the two via a logistic normal prior, such that temporal dynamics of the networks are captured" (Fu et al., 2009, p.330). This general approach to modeling time dependency in an autoregressive state-space approach is still in regular use today (e.g., Minhas et al., 2016b).

Anandkumar et al. (2014) were among the first to use tensor structures for dynamic overlapping cluster detection. They propose a learning method using spectral tensor decomposition, that is adequate for probabilistic network models with overlapping communities or *Mixed Membership Dirichlet* models. The model can learn higher-dimensional factor structures in the lower-order moments of a random variable. It directly bases on the usage of the *Mixed Membership Dirichlet* model introduced by Airoldi et al. (2008). Anandkumar et al. (2014) provided a direct link to factorization via spectral decomposition.

The *Higher Order Singular Value Decomposition* (HOSVD) produces orthogonal factors that order the latent dimensions by their overall contribution to variance, equivalently to the spectral decomposition of a matrix. We base our illustration on the article by Karatzoglou et al. (2010), which applies HOSVD in the context of collaborative filtering. Following the empirical example of the authors, let $\underline{\mathbf{Y}}$ denote a tensor of order $(N \times M \times C)$ that contains observed ratings of users on films. Here

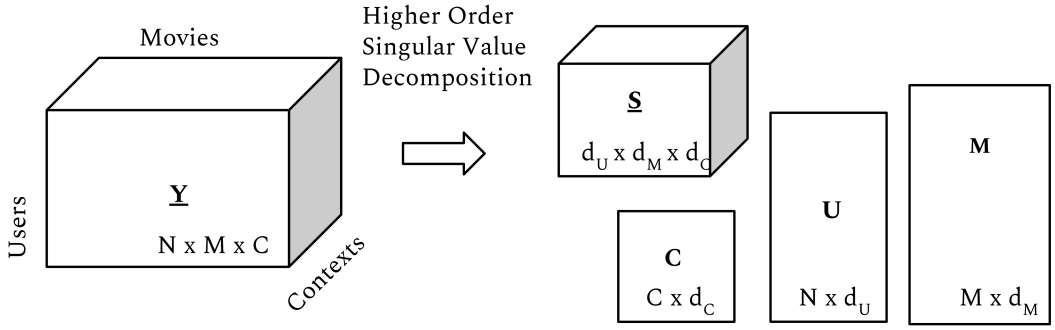

Figure 1.5: Three-dimensional tensor factorization by Higher Order Singular Value Decomposition for the example of user movie databases within multiple contexts (compare Karatzoglou et al., 2010, p.81).

N stands for the number of users, M denotes the number of films or items users have access to, and C describes the number of values associated with a contextual variable (like time or movie genre). The aim is to approximate the original tensor of observations $\underline{\mathbf{Y}}$ with a parsimonious factorial structure $\underline{\mathbf{F}}$, formally:

$$\underline{\mathbf{F}}_{nmc} = \underline{\mathbf{S}} \times_1 \mathbf{U} \times_2 \mathbf{M} \times_3 \mathbf{C} \tag{1.14}$$

that minimizes the loss between observed an approximated structure. Here, x_n ($n = 1, \ldots, 3$) denotes the so-called tensor (or Tucker) product (Cichocki et al., 2009, p.36). The factors are, like in two-dimensional models, encoded in the columns of the matrices $\mathbf{U}^{N \times d_U}$, $\mathbf{M}^{M \times d_M}$ and $\mathbf{C}^{C \times d_C}$, and $\mathbf{u}_n = \mathbf{U}_{(n \cdot)}$ is a row of the user factors, i.e. the score of observation n along all available factors (Karatzoglou et al., 2010). The central tensor $\underline{\mathbf{S}}$ is of order $(d_U \times d_M \times d_C)$, dimensions can be adjusted flexibly. Figure 1.5 illustrates the HOSVD graphically.

1.3.2 Higher Order Factorial Models

In structural prediction, the input to a prediction algorithm is a partially observed graph, while for temporal prediction, the input consists of a sequence of graphs observed at multiple time instances or within multiple time windows. The latter requires a temporal perspective, for example, via polynomial kernel based curve

fitting modeling a specific growth mechanism (Mangal et al., 2013) or via tensors (e.g., Spiegel et al., 2012). Besides temporal prediction, methodological progress in higher-order factorization had another core driver: applications in data mining on sparse network structures. In order to alleviate the sparsity of one matrix, information from a different context involving the same nodes can be helpful for prediction and pattern inference. On the other hand, treating the alternative contexts as if they represented the same form of relationship can be misguiding and not very helpful, as books and films, for example, are two contexts where use preferences exert their influence, but pooling books and films in a single set of ratings can lead to a loss of information. Therefore, recommendation models that pool together rating from several contexts without *actually pooling* the data has been on many research agendas in the last ten years (e.g., Kolda and Bader, 2009; Mørup, 2011; Spiegel et al., 2012; Hoff, 2015; Battiston et al., 2016).

The idea of solving the problem using tensor data structures itself is not new, while it has only recently become feasible for the context of large datasets. Borgatti and Everett (1992) provided seminal work on the application of blockmodeling to multimodal data such as film-actor networks with two types of nodes and multi-network data such as actor-by-actor-by-time arrays or tensors. Sidiropoulos et al. (2017) provided an overview of methodological progress in the field of tensor decomposition and its application in knowledge engineering. The authors covered a broad range of topics, including tensor factorization models, identifiability, algorithms for decomposition and use cases, "ranging from source separation to collaborative filtering, mixture and topic modeling, classification, and multilinear subspace learning" (Sidiropoulos et al., 2017, p.3551). While a comprehensive methodological overview is beyond the scope of this article, we will briefly illustrate the principle of tensor decomposition for factorization. Karatzoglou et al. (2010) emphasized the advantages of tensor factorization for recommendation tasks, as being a "generic model framework" (Karatzoglou et al., 2010, p.81) that, similarly to matrix factorization, offers fast computation and efficient optimization.

The basic idea of factorizing higher-order data structures is that multiple contexts of behavior share similar rating patterns. If there are two separate sets on book

ratings and film ratings, there are two options. On the one hand, one can pool both sets of ratings and factorize a matrix. On the other hand, it is possible to organize both matrices in the form of a *tensor* and use a tensor-based factorization. Cichocki et al. (2009) listed among the modes of analysis besides space, time and frequency also "trials, task conditions, subjects, and groups" (Cichocki et al., 2009, pp.43-44). Other examples can be different pollutants that share common factorial structures because they all depend on the distribution of industrial areas, or separate measurement facilities distributed over the country. Even if measurements are not directly comparable, information on covariance structure regarding one can be informative about the other. Information on another air pollutant can, for example, help to estimate the local factors that determine general pollution levels, when the same factors play a role in different pollutant contexts. The separate factorization of every estimated covariance matrix leads to the loss of information on the covariance structure among contexts. Multi-dimensional data structures help in retaining the cross-contextual information (Cichocki et al., 2009).

Zhang et al. (2009) partitioned the earth's surface into coherent areas with homogeneous temporal and spatial autocorrelation patterns. In their implementation, monthly variation patterns end up in the second factor, whereas a third factor represent yearly variation patterns. The corresponding global maps are recognizable in the first factor (Zhang et al., 2009). Fan et al. (2014) used tensor decomposition for analyzing people flow, assuming that spatial dynamic flows are the result of the interaction of spatial factors, for example, driven by business districts that are the target of many commuters for work, with temporal factors, when commuters arrive in the morning and leave in the evening. The authors used nonnegative tensor factorization to identify these patterns. Their analysis employs on city-wide GPS log data created by the use of smartphones. A single observation or record in the original data contains a unique ID for each check-in of a mobile phone, latitude, longitude, and time. The data are binned to the number of log-ins in a region r, day d, and time of the day t. $\underline{\mathbf{Y}} \in \mathbb{R}^{N_r \times N_t \times N_d}$ is the tensor which describes people flows, where N_r, N_t, N_d are the numbers of regions, time-slices and sample days, respectively. The authors discussed how the spectral decomposition in time can detect abnormal

population behavior after a catastrophe and analyze the flow of people that fled the Fukushima area after the nuclear catastrophe of 2011. In time factors, the outflow of people from the direct perimeter of the destroyed nuclear power plant is visible in the matrices generated by tensor decomposition.

A recent application of longitudinal network structure basing on tensor structures has been provided by Minhas et al. (2016b). The authors integrated the coevolving networks of international verbal and material interaction in a single framework. They aimed to trace two networks of $N \times N$ adjacency matrices (where N is the number of countries considered) over T periods in $V = 2$ separate contexts (i.e., verbal and material interaction). Minhas et al. (2016b) intended to understand how and why networks change over time, without declaring one of the two modes of interaction to be exogenous. Their data is stored as a $N \times N \times V \times T$ fourth-order tensor. Similar to a vector autoregressive model (VAR), the authors introduced a lagged predictor $\underline{\mathbf{Y}}_{t-1}$, and consider the V contexts to be endogenous. That is, countries that have frequent verbal cooperation (conflicts) are also more likely to have material cooperation (conflicts) and vice versa. Instead of looking at a series of two-dimensional matrices, data is organized in a series of third order tensors, where one dimension of the tensor represents context. Let $\underline{\mathbf{Y}}_t$ be a $N \times N \times V$ three-dimensional tensor, describing the network at time t. Then $\underline{\mathbf{Y}}_{t-1}$ is $\underline{\mathbf{X}}_t$ lagged by one period. $\underline{\mathbf{Y}}_{t-1}$ is used as a predictor to $\underline{\mathbf{Y}}_t$ in the regression model to trace the changes in the underlying factorial structure of the model. Using the multilinear formulation introduced in Hoff (2015), the authors obtained:

$$\underline{\mathbf{Y}}_t = \underline{\mathbf{Y}}_{t-1} \times_1 \mathbf{B}_1 \times_2 \mathbf{B}_2 \times_3 \mathbf{B}_3 + \underline{\mathbf{E}}_t. \tag{1.15}$$

$\mathbf{B}_1$ and $\mathbf{B}_2$ represent $N \times N$ matrices of regression parameters, estimated under simultaneous consideration of the V contexts that human interaction develops in. Therefore, both matrices represent the consequences of past behavior for present behavior.

$\mathbf{B}_{(1ij)}$, i.e., the i-th row in the j-th column of the matrix $\mathbf{B}_1$, represents the causal

effect of a tie from actor i to actor j along any of the V contextual modes for the likelihood that an interaction arises from actor i to actor k, one time period after. The dynamic effects of the lagged dyadic relational parameters on one another are encoded in the $V \times V$-dimensional matrix $\mathbf{B}_3$. This means that $\mathbf{B}_{3,uw}$ represent the effect of a parameter u in one network in $t-1$ on the value of a parameter w in that dyad during t. The matrix $\underline{\mathbf{E}}_t$ contains random effects (Minhas et al., 2016b).

Minhas et al. (2016b) referred to Hoff (2015) for details on the inferential implementation. While the method decomposes a three-dimensional tensor to model countries and contexts, the approach uses VAR for the integration of a fourth mode, the dynamic mode. The usage of VAR for the specification of dynamics shows a perspective on graphs that expresses stability over time. The higher the absolute values of the elements in the loading matrices, the stronger the time dependency. The relational space in the analysis is treated as endogenous, as it is the changes in the tensor structure that are of interest. Time, however, is not factorized but modeled in an autoregressive way. Time is different from other dimensions, as it has a clear causal structure: it is always the past that influences the future, not the other way around. Therefore, many models use this autoregressive formulation, which automatically implies a Markov property: Once the state of the tensor at time $t-1$ is known, the tensor state at time t is independent of the realized states in $t = 1, \ldots, t-2$. This Markov property substantially reduces the number of parameters that have to be estimated.

Next, consider geophysical applications of tensor structures, where space is considered exogenous. Henretty et al. (2017) illustrated the usage of tensor decompositions for geographic analysis with two example datasets of georeferenced multidimensional event data. The first use case seeks patterns in taxi trip data from the *Taxicab & Livery Passenger Enhancement Programs*. This dataset stems from the New York City Taxi and Limousine Commission and contains (among other features) the time of pickup, and the geographic coordinates of taxi pickup and drop-off (longitude and latitude). Henretty et al. (2017) use one week of pickup and drop-off data, containing approximately 2.5 million entries. The authors visualized affluent hours of taxi drop-offs, as well as times of high traffic and typical event locations in maps. The

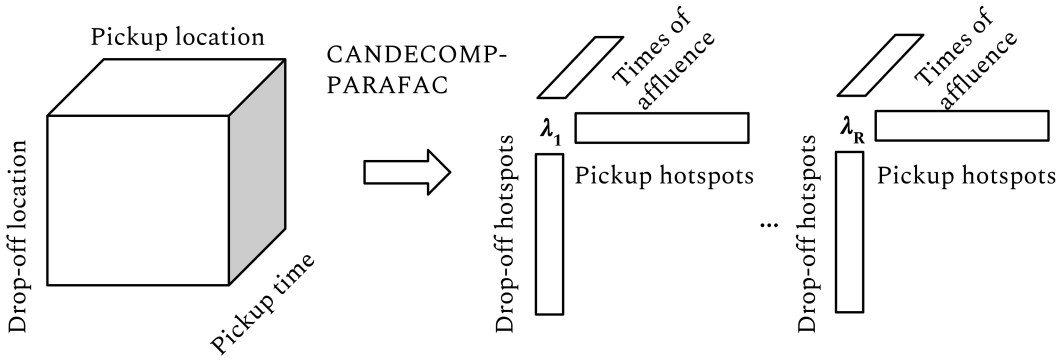

Figure 1.6: Representation of the CANDECOMP-PARAFAC decomposition. Every resulting combination represents one pattern of traffic affluence by time and location.

second dataset that the authors used contains demographic, time, and georeferenced data on traffic violations for 815.000 incidents. The dataset contains (among others) the date of the record, time of the day of the violation, the sub-agency that recorded the violation, latitude and longitude of the event, a description of the violation, and race and gender of the traffic offender. The authors applied the decomposition to a 7^{th} order tensor to analyze a dataset. Henretty et al. (2017) use a CANDECOMP-PARAFAC (CP) tensor decomposition, which is similar to HOSVD and produces orthogonal components. Like eigenvectors or singular vectors, the columns contain a score for each index of the mode. Figure 1.6 visualizes the algorithm for the first example .

The computational costs of tensor factorization were longtime prohibitive. It is not only the storage of C-modal tensor structures, which grow at order $O(n^C)$, that is, when doubling the number of observations for every dimension, the number of entries in the extended tensor equals 2^C times the number of the entries in the original tensor. As Henretty et al. (2017) explained in more detail, the precision in the reconstruction of the tensor via the outer product of the factors also decreases, and the ability of a decomposition result to accurately represent the original data decreases with the number of dimensions. Hence, the ideal rank of the decomposition of a tensor is therefore usually a lot higher than for matrix decomposition (Henretty et al., 2017). Therefore, Higher-order tensors also require more complex factor structures to represent the same relative amount of information than matrices.

1.3.3 Point-Of-Interest Recommendation

The simultaneous consideration of three and more dimensions of space at a time has opened the door for many use cases that integrate social and geographic space. At the same time that higher-dimensional factorization has become feasible, the usage of LBSNs has started producing a lot of geotagged, commercially valuable data. Recommendations with LBSN data can process both the social network of a user, as well as her physical surroundingsi. For example, the physical distance to a possible *Point Of Interest* (POI)[2] can be as important as the question whether many friends of a user have visited the POI.

Cheng et al. (2012) argued that t four general influence factors drive the choice of the next POI of individual i. *First*, data on check-ins only contains positive entries. That is, the absence of a tie between a user and a POI does not mean that the user would not visit the place. Usually, LBSN data is very sparse. *Second*, POIs are usually clustered around some centers, which might be the workplace and private home, as well as the nearby city center. Usually, the probability of visiting a POI follows a Gaussian distribution around these centers. *Third*, the probability of visiting a POI declines concentrically with distance to the nearest center, leveling out to a probability of zero when the POI is too far away from the centers a person i usually frequents. The *fourth* characteristic consists in *friendship influence*. For the dataset that Cheng et al. (2012) used, the mean overlap of visited POIs by a user and the POIs visited by his/her friends is about 9.6%, a ratio that is approved by other authors (e.g., Cho et al., 2011). The social embedding's influence seems limited but might help to reduce cold start problems[3], as shown by Gao et al. (2012).

He et al. (2016) categorized POI recommendation into four branches, and name examples for every category: *Time-aware*, *geographical influence enhanced*, *content-aware* and *social influence enhanced* POI recommendation. To predict which location people might be interested in going to next, models choose from the available dimensions of dependency. A time-aware analysis seeks to model periodic patterns

[2]POIs are potentially interesting locations like bars, monuments, or restaurants that a person might "check-in" to.

[3]Cold start problems arise when only very little information on a person is given, which substantially reduces the performance of collaborative filtering algorithms.

like the time of the day of user habits. Geographic analysis recognizes the restrictions imposed on a person by being somewhere and not elsewhere and, therefore, in the reach of some POI. Content-aware recommendation uses textual context like Twitter messages to issue recommendations. Social influence models assume that people like doing what their peers like doing and leverage the social embedding of a person.

Authors such as Yuan et al. (2013) and Yao et al. (2015) pointed out a fifth mode, emphasizing the fact that check-in data is highly periodical, and different types of POI show high dependence on different temporal factors. For example, factorial methods can identify lunch-time or nightclub-time from the data without the need to explicitly specify predictors for this behavior. Zhao et al. (2016) circumscribed user mobility in LBSNs as showing "distinct temporal features, summarized as periodicity, consecutiveness, and non-uniformness" (Zhao et al., 2016, p.450). Periodicity describes phenomena like the daily visit to a café before going to work. In order to analyze the three patterns together, the authors suggested an *Aggregated Temporal Tensor Factorization* model for POI recommendation. The analysis concentrates on the transition probability to go from one category of location to another category.

So what techniques are best suited for this context? As early pioneers of the topic, Ye et al. (2010) relied on matrix-based collaborative recommendation for POI recommendation. The authors used heuristics for measuring the strength of social and geospatial ties among users and their visited locations. Little later, Cheng et al. (2012) used matrix factorization in combination with geographical and social influence for POI recommendation in LBSNs. The authors included the role of geographic proximity by modeling a user's probability to next check-in a specific POI with a multi-center Gaussian model. Social information is then fused into a generalized matrix factorization framework. Zhang and Chow (2013) chose a similar approach, developing a recommendation model by fusing Kernel density estimation into the matrix factorization framework.

A joint shortcoming of these approaches is that matrices can only represent two dimensions at a time. Additional information have to be included in additional model terms or heuristics (like cut-off rules). Yao et al. (2015) were among the first

to use a tensor-based approach to combine geographical and social information for POI recommendation. The authors proposed a collaborative filtering approach to POI recommendation, which uses a tensor factorization on a high-order tensor.

He et al. (2016) used a dynamic algorithm to predict *Points Of Interest* (POI) based on geographic proximity, individual characteristics/interest profiles, and dynamic information concerning previous choices. Central to their model is the observation that, under different contextual scenarios, humans exhibit distinct mobility patterns. For example, people "regularly stop by coffee stalls [...] to grab a cup of coffee on their way to work in the morning, which can be explained as a periodic transition pattern from coffee shop to workplace on weekday morning" (He et al., 2016, p.137). Hence, assuming that some latent behavior patterns govern individual mobility, the authors jointly inferred the pattern distribution and pattern-level transition probabilities that indicate the likeliness to go from one to another place.

Adapting He et al. (2016), let $y_{u,i,l}$ denote historical transitions between locations i and l for a particular user u, where $i, l \in \{1, \ldots, L\}$ and $u \in \{1, \ldots U\}$. $y_{u,i,l} = 1$ if a transition can be observed and $y_{u,i,l} = 0$[4] otherwise. The data can be organized as a tensor $\underline{\mathbf{Y}}$ of order $(U \times L \times L)$. Now, the aim is to estimate transition probabilities

$$x_{u,i,l} = P(y_{u,i,l} = 1) \tag{1.16}$$

that best explain historical transitions and can be used for POI recommendation in the next time step. For that purpose, the authors resorted to a vector of contextual features $\mathbf{c} = (c_1, \ldots, c_F)$ which include previous location, time of day, weekday, previous location's category, and so on. Then, if $s,\ s \in \{1, \ldots, S\}$ describes a context-dependent latent behavior pattern, the transition probabilities to be estimated can be expressed as

$$P(y_{u,i,l} = 1) = \sum_{s=1}^{S} P(y_{u,i,l} = 1 | s, \mathbf{c}) \cdot P(s|\mathbf{c}), \tag{1.17}$$

[4]Note that He et al. (2016) did not explicitly state the alternative value.

that is, the sum of the joint probabilities concerning transitions and pattern-level, conditioned on the context. Thus, the overall goal of estimating (unconditional) transition probabilities is to be replaced by the joint estimation of pattern distribution and pattern-sensitive transition probabilities.
As the elements $\hat{x}_{u,i,l} = P(y_{u,i,l}|s, \mathbf{c})$ can, once again, be organized as a tensor $\underline{\hat{X}}^s$ of order $(U \times L \times L)$, it is possible to apply the so-called *Pairwise Interaction Tensor Factorization* (PITF):

$$\hat{x}_{u,i,l} = (\mathbf{u}^s_{u,l})^T \mathbf{l}_{l,u} + (\mathbf{l}^s_{l,i})^T \mathbf{i}_{i,l} + (\mathbf{u}^s_{u,i})^T \mathbf{i}_{i,u} + \rho d^{-1}_{i,l}. \tag{1.18}$$

The single vectors in equation (1.18) describe the latent factors of User u in interaction with current location l and next location i ($\mathbf{u}_{u,l}$ and $\mathbf{u}_{i,u}$, respectively), of present location l in interaction with user u and next location i ($\mathbf{l}_{l,u}$ and $\mathbf{l}_{l,i}$, respectively), and of next location i with present location l and user u ($\mathbf{i}_{i,l}$ and $\mathbf{i}_{i,u}$, respectively). The interpretation of this equation is quite similar to the interpretation of equation (1.12). With reference to Rendle et al. (2009), the authors argued that $(\mathbf{u}_{u,i})^T \mathbf{i}_{i,u}$ can be removed from the model since it doesn't affect the ranking. $\rho d^{-1}_{i,l}$ expresses spatial preference meaning that the probability to visit a POI that is $d_{i,l}$ kilometers away from the current position of user u declines exponentially with distance. Since the locations previously visited a relegated to the context and the current position of the user is treated as exogenous, the transition probabilities to be calculated in He et al. (2016) exhibit the Markov property.

1.4 Discussion

1.4.1 Geographical and Social Network Space

From a mathematical perspective, geographical space and social network space have many things in common. Spatial dependency in geographic data is often encoded in *neighborhood matrices* (e.g., Wang and Wall, 2003). The neighborhood matrix is very similar to the *adjacency matrix* of a graph, which is also often binary and

indicates whether there is a connection between two nodes or not (e.g., Hoff, 2009). In geographical information science, it is sometimes even better to define physical distance in terms of networks. Road networks may provide a more realistic operationalization of neighborhood than geophysical distance: Road networks link one house or town to another and determine how easy it is to get to a different place. Therefore, using such measures of distance can sometimes explain human interaction better than physical distance (Borruso, 2008). Network neighbors tend to show similar or opposite characteristics, depending on whether there is a positive or negative dependency between them. For example, cooperating companies in a collaboration network tend to positively influence one another, while competitors at times show different performance development and negative dependency.

Latent variable models focus on modeling dependencies that arise in spatial data. A latent variable is an unobserved construct causing dependency, which can, for example, be a local geographic phenomenon that causes heterogeneity in the concentration of air pollutants. For example, a polluting production site may be missing in the available data, which causes temporal dependency in air pollution by emitting pollutants at a specific time of the day, and geographical dependency by omitting pollutants only at its geophysical location. In dynamic space, latent variables can describe a general downturn in the economy that affects single countries in varying intensity. In social space, latent variables can describe a common unobserved hobby, which raises the propensity that two people show interest in one another. In graphical data or network data, they propose possible "friends", or products that might match people's interests, basing on information about the network embedding. In the analysis of geophysical data, factorial models identify local "interesting patterns" with a limited variance-covariance structure.

In this article, we ask whether the class of latent factor models contains any possibility to assess such a perspective, that is, to what extent algorithms explicitly involve their own influence on homogeneity. We show in the discussion that often, geographical space is exogenous to the analysis when referring to geospatial analysis. Exogenous space means that the position of an individual is fixed and does not change depending on influence factors. Endogenous means that the relative

position of a person changes due to some exogenous or endogenous influence factors. For example, when a person changes her position in a city to a borrow where she finds more recommended restaurants, her geospatial position is endogenous to the analysis. When algorithms start influencing the structures they try to describe, then model elements that are assumed to be exogenous become endogenous. If, over time, people start befriending only people from their recommendation lists, then the "reason" of proximity lies in the algorithm itself. A particular focus of our investigation is on the differentiation of virtual and geographical space. Both dimensions of space are essential for understanding filter bubbles, as the defendants of the filter bubble hypothesis emphasize that the algorithms do not only *describe* social space, as they might seem to at first sight. The analysis shows that many models treat social and physical space similarly as exogenous in the aim to optimize a short-term predictive performance. Indeed, it may also be the case that physical proximity is the result of a similarity in the first place – people may, for example, decide to move to a neighborhood of culturally similar people.

1.4.2 Big Data and Bias

Many applications involving latent variable models are related to the aim to generate value from the high-dimensional and complex data structure that evolves from modern communication systems. The high frequency, precise geospatial localization, and feature dimensionality of the data invite to use concepts that were historically used in well-curated small datasets in these contexts. The filter bubble discussion is related to the negative consequences that may arise from such pragmatism. When the risks related to the algorithms are not properly accounted for, adverse effects such as discrimination can arise.

If human mindsets were independent of changes in geography and networks, then filters would merely *describe*. For statistical analysis and descriptive analytics, assumptions are needed that restrict the many interactions between the dimensions, to keep computation feasible and models identified. For a short-term investigation, it might, for example, be reasonable to assume that firms cannot change their geographic situation. In the long term, however, they can found new dependencies

and optimize their geographic situation. A changed geographic situation may then again influence the social embedding of the firm among strategic partners. While exploratory factorization methods often approach data analytics as though they were on the quest for exogenous, deterministic underlying patterns, they risk becoming the drivers of self-fulfilling prophecies, obeying to the aim of providing correct propositions.

In statistical language, unobserved patterns that cause clustering and observed heterogeneity translate into models involving latent variables and factors. Social structures and geographical phenomena naturally P dependency between people that interact or live nearby. Therefore, cross-sectional dependency is an ever-present phenomenon in data on human behavior or spatially referenced data. Any form of dependency between individual observations, either in time or in space, represents a violation of classical assumptions of independence and exchangeability (Hoover, 1982; Aldous, 1985), which are necessary for the application of many models. Typical phenomena that lead to spatial dependency are common environmental influences (like the presence of a polluting fabrication hall), talking to one's neighbors, reciprocity (people are more likely to consider people friends that would themselves consider the other person as a friend), or heterogeneity in combination with homophily (people like having contact with similar people) (Hoff et al., 2002; Hoff, 2008, 2009; Pesaran and Tosetti, 2011).

Statistical model treat this unobserved dependency, which manifests in observed correlation, in two different ways. On the one hand, the dependency can be a problem that impedes the unbiased inference of parameters in models. In this case, dependency must be "dealt with". For example, an interactive factorial structure can be included in the error term of a model (e.g., Bai, 2009; Pesaran and Tosetti, 2011). On the other hand, spatial dependency can also be seen a source of information worth mining for recommendation or prediction: Patterns of correlation provide valuable information on unobserved spatial signals, interaction patterns, and preferences (e.g., Hoff, 2009; Demšar et al., 2013). Exploratory latent variable approaches assess dependency and similarity structures in a scalable and efficient way, as some methods for factor extraction are of low computational cost. Therefore, they also allow

exploring the massive datasets that virtual interaction creates every day and to use them for pattern identification and link prediction (Hoff et al., 2002; Koren et al., 2009; Demšar et al., 2013; Minhas et al., 2016a). Moreover, factorization methods allow combining implicit with explicit spatial information and observable covariates (Menon and Elkan, 2011). Exploratory latent variable approaches moreover bear the advantage of being independent of explicitly denominated hypotheses on structures and relationships within the data. In other contexts, when statistical inference is in the center of attention, complex Bayesian model specifications have been proposed. These probabilistic models, based on the principles of prior belief and posterior likelihood, provide more precise estimations of spatial effects.

1.5 Conclusions

This article provided an overview of latent variable models and the role that physical and virtual space play therein. The discussed models seek to either compensate for or mine the presence of dependency structures via the inclusion of latent variables of either categorical (group) or numeric (factor) nature. Recent methodological progress within the field concentrates on three- and higher-dimensional models. These models explore the simultaneous interplay of several modes in data. While this used to be computationally infeasible in the past, modern computers and algorithms can resolve complex tensor factorization.

This article describes the way that social and geographical space are modeled in latent variable models and related recommender and prediction algorithms. Our first focus was on a description of the interplay of different dimensions of space in factor-based models. In factor models, mindsets, convictions, taste, or local air pollution conditions are assessed via feature space, while we denoted social ties as social network space. Both dimensions can be combined with geographical space and time. Moreover, in higher-order factorial models, cross-contextual factorization can be used to find similar patterns across different settings. To illustrate the concepts of exogenous and endogenous space, we chose to discuss geographical space under the aspect of being fixed, and network space under the aspect of being in constant

change and being endogenous to a recommender model. While it is possible to model both structures also in the both ways, we illustrated the difference between finding signals in fixed space and identifying patterns in the spatial structure itself.

The discussion also showed that while two-dimensional methods were limited to the factorization of two dimensions at a time, more recent tensor-based approaches can model a complex dynamic interplay of several perspectives on space. However, tensor decomposition is computationally more demanding. Despite the fact that higher-order tensor factorization is possible (remember the seven-dimensional example of Henretty et al., 2017), most approaches that go beyond the usage of three dimensions at a time use a three-dimensional tensor combined with a vector autoregressive element to trace changes along time and to make a specific time-related prediction (e.g., Minhas et al., 2016b; He et al., 2016). Modeling time in an autoregressive element is an obvious choice because time is unidirectional, whereas geographical space is at least two-directional and network space can be multidirectional.

The focus of higher-dimensional factor models has so far been mostly on predictive performance. As long as the patterns and their interpretation are not part of the model objectives and do not have any direct policy implications, a possible misunderstanding of the nature of the estimated factors has no direct consequences. However, once it comes to the interpretation of such results, the same risks arise as in the two-dimensional context. In virtual networks, distance is subject to constant change, as the distance of two people changes whenever a proposition of adding someone as a contact is accepted. This phenomenon is more accentuated than in "traditional" social environments, which were dominated by physical proximity.

To avoid adverse effects of algorithms on society, explainability in algorithms is very important. Zhang and Chen (2018) discuss recent progress in creating explainable recommender systems, where the analysis of in- and output can provide crucial insights into model bias. The unreflected use of latent-variable models and factorization can lead to a self-fulfilling prophecy - people consume what they are confronted with, and believe what their virtual community believes. The more physical exchange of opinions is replaced by virtual exchange with groups of people tailored to one's preferences, the more readily people feel confirmed in their own convictions.

While many aspects of statistical models regarding the self-reflection have been discussed, the question of strategy behind the algorithms is still open. If the aim is to "be right" with one's recommendations, then a tendency towards homogenization is warranted to make behavior more predictable. Such objectives might, however, also be warranted by the users of an algorithm. Even *if* factorial structures become more clear-cut over time, it can be argued that preferences had been clear-cut and clustered before and that the recommender algorithms only better succeed in describing those preferences.

The enhanced understanding of assumptions and consequences of every single model is essential for the discussion of changes in society that come along with the massive commercial exporation of correlations in a data-driven industry. Descriptive approaches like the ones in the work of Nguyen et al. (2014) and Bakshy et al. (2015) can describe the diversity people are exposed to. Our approach has shown that many recommendation algorithms ignore their own role in changes of the patterns they seek to describe. A complementary step would be to conduct experiments that show that people do indeed change when they use recommender algorithms.

1.6 References

Abadie, A., Diamond, A., Hainmueller, J., 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association 105, 493–505. doi:10.1198/jasa.2009.ap08746.

Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P., 2008. Mixed membership stochastic blockmodels. Journal of Machine Learning Research 9, 1981–2014.

Aldous, D.J., 1985. Exchangeability and related topics, in: École d'Été de Probabilités de Saint-Flour XIII –1983. Springer, Berlin, Heidelberg, GER, 1–198.

Anandkumar, A., Ge, R., Hsu, D.J., Kakade, S.M., 2014. A tensor approach to learning mixed membership community models. Journal of Machine Learning Research 15, 2239–2312.

Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica 77, 1229–1279. doi:10.3982/ECTA6135.

Bailey, T.C., Krzanowski, W.J., 2000. Extensions to spatial factor methods with an illustration in geochemistry. Mathematical Geology 32, 657–682. doi:10.1023/A:1007589505425.

Bakshy, E., Messing, S., Adamic, L.A., 2015. Exposure to ideologically diverse news and opinion on Facebook. Science 348, 1130–1132. doi:10.1126/science.aaa1160.

Battiston, F., Iacovacci, J., Nicosia, V., Bianconi, G., Latora, V., 2016. Emergence of multiplex communities in collaboration networks. PLOS ONE 11, 1–15. doi:10.1371/journal.pone.0147451.

Berry, B.J.L., 1971. Introduction: The logic and limitations of comparative factorial ecology. Economic Geography 47, 209–219. doi:10.2307/143204.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Borgatti, S.P., Everett, M.G., 1992. Regular blockmodels of multiway, multimode matrices. Social Networks 14, 91–120. doi:10.1016/0378-8733(92)90015-y.

Borruso, G., 2008. Network density estimation: A GIS approach for analysing point patterns in a network space. Transactions in GIS 12, 377–402. doi:10.1111/j.1467-9671.2008.01107.x.

Cheng, C., Yang, H., King, I., Lyu, M.R., 2012. Fused matrix factorization with geographical and social influence in location-based social networks, in: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 17–23.

Cho, E., Myers, S.A., Leskovec, J., 2011. Friendship and mobility: User movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. 1082–1090. doi:10.1145/2020408.2020579.

Christensen, W.F., Amemiya, Y., 2001. Generalized shifted-factor analysis method for multivariate geo-referenced data. Mathematical Geology 33, 801–824. doi:10.1023/A:1010998730645.

Christensen, W.F., Amemiya, Y., 2002. Latent variable analysis of multivariate spatial data. Journal of the American Statistical Association 97, 302–317. doi:10.1198/016214502753479437.

Christensen, W.F., Amemiya, Y., 2003. Modeling and prediction for multivariate spatial factor analysis. Journal of Statistical Planning and Inference 115, 543–564. doi:10.1016/s0378-3758(02)00173-8.

Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.i., 2009. Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons, Chichester, UK.

Cressie, N., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. Wiley Series in Probability and Statistics, John Wiley & Sons, New Jersey, USA.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S., McLoone, S., 2013. Principal component analysis on spatial data: an overview. Annals of the Association of American Geographers 103, 106–128. doi:10.1080/00045608.2012.689236.

Dommenget, D., Latif, M., 2002. A cautionary note on the interpretation of EOFs. Journal of Climate 15, 216–225. doi:10.1175/1520-0442(2002)015<0216:acnoti>2.0.co;2.

Fan, Z., Song, X., Shibasaki, R., 2014. CitySpectrum: A non-negative tensor factorization approach, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), 213–223. doi:10.1145/2632048.2636073.

Ferligoj, A., Doreian, P., Batagelj, V., 2011. Positions and roles, in: Scott and Carrington (2011). Chapter 29. 434–446. doi:10.4135/9781446294413.n21.

Fu, W., Song, L., Xing, E.P., 2009. Dynamic mixed membership blockmodel for evolving networks, in: Proceedings of the 26th Annual International Conference on Machine Learning (ICML), 329–336. doi:10.1145/1553374.1553416.

Gao, H., Tang, J., Liu, H., 2012. gSCorr: modeling geo-social correlations for new check-ins on location-based social networks, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), 1582–1586. doi:10.1145/2396761.2398477.

Goldberg, K., Roeder, T., Gupta, D., Perkins, C., 2001. Eigentaste: A constant time collaborative filtering algorithm. Information Retrieval 4, 133–151. doi:10.1023/A:1011419012209.

Handcock, M.S., Raftery, A.E., Tantrum, J.M., 2007. Model-based clustering for social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society) 170, 301–354. doi:10.1111/j.1467-985x.2007.00471.x.

He, J., Li, X., Liao, L., Song, D., Cheung, W.K., 2016. Inferring a personalized next point-of-interest recommendation model with latent behavior patterns, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 137–143.

Henretty, T., Baskaran, M., Ezick, J., Bruns-Smith, D., Simon, T.A., 2017. A quantitative and qualitative analysis of tensor decompositions on spatiotemporal data, in: 2017 IEEE High Performance Extreme Computing Conference (HPEC), 1–7. doi:10.1109/HPEC.2017.8091028.

Hoff, P.D., 2008. Modeling homophily and stochastic equivalence in symmetric relational data, in: Platt, J.C., Koller, D., Singer, Y., Roweis, S.T. (Eds.), Advances in Neural Information Processing Systems. Volume 20, 657–664.

Hoff, P.D., 2009. Multiplicative latent factor models for description and prediction of social networks. Computational and Mathematical Organization Theory 15, 261–272. doi:10.1007/s10588-008-9040-4.

Hoff, P.D., 2015. Multilinear tensor regression for longitudinal relational data. The Annals of Applied Statistics 9, 1169–1193. doi:10.1214/15-AOAS839.

Hoff, P.D., Raftery, A.E., Handcock, M.S., 2002. Latent space approaches to social network analysis. Journal of the American Statistical Association 97, 1090–1098. doi:10.1198/016214502388618906.

Hogan, J., Tchernis, R., 2004. Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. Journal of the American Statistical Association 99, 314–324. doi:10.1198/016214504000000296.

Hoover, D.N., 1982. Row-column exchangeability and a generalized model for probability, in: Proceedings of the International Conference on Exchangeability in Probability and Statistics, North-Holland. 281–291.

Johnston, R., Pattie, C., 2011. Social networks, geography and neighbourhood effects, in: Scott and Carrington (2011). Chapter 21. 301–311. doi:10.4135/9781446294413.n21.

Jolliffe, I.T., 1986. Principal component analysis and factor analysis, in: Principal Component Analysis. Springer, New York, NY, USA, 115–128. doi:10.1007/978-1-4757-1904-8_7.

Jolliffe, I.T., 1987. Rotation of principal components: some comments. Journal of Climatology 7, 507–510.

Karatzoglou, A., Amatriain, X., Baltrunas, L., Oliver, N., 2010. Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering, in: Proceedings of the fourth ACM Conference on Recommender Systems (RecSys), 79–86. doi:10.1145/1864708.1864727.

Kolaczyk, E.D., Csárdi, G., 2014. Statistical Analysis of Network Data with R. Volume 65. Springer, New York, NY, USA.

Kolda, T.G., Bader, B.W., 2009. Tensor decompositions and applications. SIAM review 51, 455–500. doi:10.1137/07070111x.

Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. Computer 42, 30–37. doi:10.1109/MC.2009.263.

Kunegis, J., Fay, D., Bauckhage, C., 2010. Network growth and the spectral evolution model, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), 739–748. doi:10.1145/1871437.1871533.

Loehlin, J.C., 2004. Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis. Taylor & Francis, New York, USA.

Lopes, H.F., Gamerman, D., Salazar, E., 2011. Generalized spatial dynamic factor models. Computational Statistics & Data Analysis 55, 1319–1330. doi:10.1016/j.csda.2010.09.020.

Mangal, D., Sett, N., Singh, S.R., Nandi, S., 2013. Link prediction on evolving social network using spectral analysis, in: IEEE International Conference on Advanced Networks and Telecommuncations Systems, 2013, 1–6. doi:10.1109/ants.2013.6802867.

Manovich, L., 2018. The science of culture? social computing, digital humanities and cultural analytics. URL: osf.io/preprints/socarxiv/b2y79, doi:10.22148/16.004.

Menon, A.K., Elkan, C., 2011. Link prediction via matrix factorization, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 437–452. doi:10.1007/978-3-642-23783-6_28.

Minhas, S., Hoff, P.D., Ward, M.D., 2016a. Inferential Approaches for Network Analyses: AMEN for Latent Factor Models. Technical Report. Cornell University. ArXiv preprint arXiv:1611.00460.

Minhas, S., Hoff, P.D., Ward, M.D., 2016b. A new approach to analyzing coevolving longitudinal networks in international relations. Journal of Peace Research 53, 491–505. doi:10.1177/0022343316630783.

Mørup, M., 2011. Applications of tensor (multiway array) factorizations and decompositions in data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1, 24–40. doi:10.1002/widm.1.

Nguyen, T.T., Hui, P.M., Harper, F.M., Terveen, L., Konstan, J.A., 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity, in: Proceedings of the 23rd International Conference on World Wide Web (WWW), ACM. 677–686. doi:10.1145/2566486.2568012.

Onnela, J.P., Arbesman, S., González, M.C., Barabási, A.L., Christakis, N.A., 2011. Geographic constraints on social network groups. PLOS ONE 6, 1–7. doi:10.1371/journal.pone.0016939.

Pariser, E., 2012. The filter bubble: How the new personalized web is changing what we read and how we think. Penguin Books, London, UK.

Pesaran, M.H., Tosetti, E., 2011. Large panels with common factors and spatial correlation. Journal of Econometrics 161, 182–202. doi:10.1016/j.jeconom.2010.12.003.

Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L., 2009. BPR: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI), 452–461. URL: http://dl.acm.org/citation.cfm?id=1795114.1795167.

Richman, M.B., 1986. Rotation of principal components. Journal of Climatology 6, 293–335. doi:10.1002/joc.3370060305.

Salakhutdinov, R., Mnih, A., 2007. Probabilistic matrix factorization, in: Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS). NIPS'07, 1257–1264.

Salakhutdinov, R., Mnih, A., 2008. Bayesian probabilistic matrix factorization using markov chain monte carlo, in: Proceedings of the 25th International Conference on Machine Learning (ICML), ACM. 880–887. doi:10.1145/1390156.1390267.

Scott, J., Carrington, P. (Eds.), 2011. The SAGE Handbook of Social Network Analysis. SAGE Publications, London, UK.

Sidiropoulos, N.D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E.E., Faloutsos, C., 2017. Tensor decomposition for signal processing and machine learning. IEEE Transactions on Signal Processing 65, 3551–3582. doi:10.1109/tsp.2017.2690524.

Snijders, T.A., Nowicki, K., 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. Journal of Classification 14, 75–100. doi:10.1007/s003579900004.

Spiegel, S., Clausen, J., Albayrak, S., Kunegis, J., 2012. Link prediction on evolving data using tensor factorization, in: New Frontiers in Applied Data Mining: PAKDD 2011 International Workshops, Springer. 100–110. doi:10.1007/978-3-642-28320-8_9.

Switzer, P., Green, A.A., 1984. Min/max autocorrelation factors for multivariate spatial imagery. Technical Report 6. Stanford University. URL: https://ci.nii.ac.jp/naid/10017502204/en/.

von Storch, H., 1999a. Misuses of statistical analysis in climate research, in: Analysis of Climate Variability. Springer, Berlin, Heidelberg, GER, 11–26. doi:10.1007/978-3-662-03744-7_2.

von Storch, H., 1999b. Spatial patterns: EOFs and CCA, in: Analysis of Climate Variability. Springer, Berlin, Heidelberg, GER, 231–263. doi:10.1007/978-3-662-03744-7_13.

Wang, F., Wall, M.M., 2003. Generalized common spatial factor model. Biostatistics 4, 569–582. doi:10.1093/biostatistics/4.4.569.

Wang, P., Robins, G., Pattison, P., Lazega, E., 2013. Exponential random graph models for multilevel networks. Social Networks 35, 96–115. doi:10.1016/j.socnet.2013.01.004.

Wang, Y.J., Wong, G.Y., 1987. Stochastic blockmodels for directed graphs. Journal of the American Statistical Association 82, 8–19. doi:10.1080/01621459.1987.10478385.

Yao, L., Sheng, Q.Z., Qin, Y., Wang, X., Shemshadi, A., He, Q., 2015. Context-aware point-of-interest recommendation using tensor factorization with social regularization, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1007–1010. doi:10.1145/2766462.2767794.

Ye, M., Yin, P., Lee, W.C., 2010. Location recommendation for location-based social networks, in: Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 458–461. doi:10.1145/1869790.1869861.

Yuan, Q., Cong, G., Ma, Z., Sun, A., Thalmann, N.M., 2013. Time-aware point-of-interest recommendation, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, 363–372. doi:10.1145/2484028.2484030.

Zhang, J.D., Chow, C.Y., 2013. iGSLR: personalized geo-social location recommendation: a kernel density estimation approach, in: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 334–343. doi:10.1145/2525314.2525339.

Zhang, Q., Berry, M.W., Lamb, B.T., Samuel, T., 2009. A parallel nonnegative tensor factorization algorithm for mining global climate data, in: Proceedings of the 9th International Conference on Computational Science (CORES), Springer. 405–415. doi:10.1007/978-3-642-01973-9_45.

Zhang, Y., Chen, X., 2018. Explainable recommendation: A survey and new perspectives. CoRR abs/1804.11192.

Zhao, S., Lyu, M.R., King, I., 2016. Aggregated temporal tensor factorization model for point-of-interest recommendation, in: Proceedings of the 23rd International Conference on Neural Information Processing (ICONIP), Part III, Springer. 450–458. doi:10.1007/978-3-319-46675-0_49.

Zhu, F., Chen, G., Heng, P.A., 2016. A Bayesian nonparametric approach to dynamic dyadic data prediction, in: 16th International Conference on Data Mining (ICDM), IEEE. 729–738. doi:10.1109/icdm.2016.0084.

Zwiers, F.W., von Storch, H., 2004. On the role of statistics in climate research. International Journal of Climatology 24, 665–680. doi:10.1002/joc.1027.

Chapter 2

Spectral Stability in Open-Source Software Developer Networks

With special thanks to Prof. Dr. Sven Apel and Thomas Bock, Faculty of Computer Science and Mathematics, Chair of Software Engineering I, University of Passau

Chapter Abstract

Open-source software development projects have a reputation for being anarchic compared to commercial projects. Previous research has found coordination efforts to arise ad hoc around programming tasks, which implies that only seldom long-term stable structures arise. However, the methodology to describe and quantify organizational stability and congruence of coordination and cooperation in software development lacks intuitive tools for description and visualization, such that it is hard to reject or approve this opinion. We complement existing approaches with a new model to investigate the stability of unobserved group structures in collaboration networks. To this aim, we combine a static, inferential network model (called AMEN) with a dynamic, predictive network model (called *spectral growth model*). The resulting model can both track the dynamics of organizational structures and check the congruence of coordination efforts and needs. We apply our model to the communication and cooperation among developers in three substantial and popular open-source projects, `QEMU`, `OpenSSL`, and `BusyBox`. We find that the congruence of email communication and collaboration on source code is relatively weak and that there is only minor evidence for growing stability, be it over time or with increasing size of the network.

2.1 Introduction

Open-source software (OSS) development projects often depend on a relatively small group of developers, who are accountable for a large share of code contributions and coordination efforts. These *core developers* are supported by a large number of *peripheral* contributors, who invest less time and effort individually, but together add a substantial part of the value in open-source software creation (Setia et al., 2012; Crowston and Shamshurin, 2017; Joblin et al., 2017). Together, the core and peripheral contributors form open-source *communities*, that create significant contributions like the Linux Kernel. Within the communities, *sub-groups* or *sub-communities* arise that work on a topic or task for a limited time only (Bird, 2011). Decades of version control and email list data provide detailed insights into the dynamics of social bonds and operational structures in this highly volatile working environment. OSS projects follow organizational rules that differ from those of conventional non-disclosed software with regard to many different aspects, such as contributor motivation (Shah, 2006), communication (Guzzi et al., 2013), and professionalization of programmers (Homscheid et al., 2015). It is, as one of many aspects, highly relevant to understand the operational structure of open-source projects in terms of collaboration and communication to diagnose sources of problems and threats to long-term stability and success.

The stability of the organizational structure of a project can be assessed via latent-factor-based models, as stability of social relations gives rise to observable correlations across network ties. Methodology to assess stability in multi-modal social networks, however, is not well established. For example, many macroscopic measures rely on a single descriptive measure of the network, which is inadequate for predicting the future development of the network. Other methods aggregate too heavily and do not show the dynamics of a changing network. Therefore, this article has both an empirical and a methodological aim. On the one hand, we analyze the degree of stability that arises in collaboration and communication habits of OSS projects, and whether this stability depends on project scale. We seek to find out whether there are stable relationships in open-source communication networks that

go beyond the day-to-day ad hoc collaboration. On the other hand, we contribute to methodological progress by developing a combination of two models from factorial network analysis, the *spectral growth model*, proposed by Kunegis et al. (2010), and the *Additive and Multiplicative Effects model*, proposed by Hoff (2009). This combination of a flexible cross-sectional framework based on Monte Carlo Estimation strategies, with the dynamic perspective of the *spectral growth* model, provides insights into the dynamics of OSS developer cooperation and the stability of the organizational structure and social bonds. For example, our measures reflect how stable modular structures are. By validating the predictive performance of our new approach, we seek to assess the strength of short-term relationships between communication and collaboration. For this purpose, we use communication as a predictor of collaboration. For our empirical investigation, we use version control data on programming tasks and email communication meta-data.

The rest of the article is structured as follows: In Section 2.2, we line out the methodological scope of our research context, and we motivate an integrated perspective on network topology and dynamics in OSS networks. Having defined core objectives of the model development and general research questions, we detail the *additive and multiplicative effects* model by Hoff (2009) and the *spectral growth model* by Kunegis et al. (2010) (Section 2.3). Both propose partial but complementary solutions to the investigation of developer cooperation behavior and the investigation of our research questions. Adaptations of the models allow us to infer a project's cooperation structure, and to track the stability of this structure over time. For this purpose, we use measures of similarity of matrices and predictive performance of stability-based network prediction methods. Moreover, our adaptations model the interplay of developer communication and collaboration. Subsequently (Section 2.4), we introduce our empirical data and data operationalization. We use communication and collaboration data on three OSS projects, `QEMU`, `OpenSSL` and `BusyBox`, retrieved from email lists and the GitHub version control system. We provide empirical results at the end of Section 2.4. Our results speak in favor of a weak relationship between communication and collaboration on a three-month time window basis. Moreover, we find weak evidence for increases in pattern stability over time or with a growing number

of programmers that participate in a project. Against our expectations, short-term collaboration seems to be no good predictor of communication, implying that coordination efforts do not respond to coordination needs in time. Section 2.6 evaluates the empirical results concerning our research questions and provides an outlook on the future scope of research and possible extensions; Section 2.7 concludes.

2.2 Dependency in Open-Source Programmer Networks

2.2.1 Organizational Structure and Collaboration

Open-source software (OSS) development is an early pioneer market concerning modern working structures: it is digital, distributed around the globe, and characterized by a high in- and outflow of workforce. These phenomena nowadays describe the working conditions of an ever-growing number of people. Interestingly, it is also highly transparent and very well documented. Therefore, it can provide informative insights into the organizational structure of such working environments, which have a somewhat anarchic reputation (e.g., Bird et al., 2008). Version control systems support detailed documentation on programming progress, and version control data are usually publicly available for open-source projects. As Draheim and Pekacki (2003) and Fischer et al. (2003) argued, the importance of version control data for the optimization of software development processes was recognized in the early days of the development of version control systems. Therefore, the usability of the data for process-based analysis always influenced the the design of version control systems, and version control data provide an excellent source of information on the projects' dynamics.

In 2008, Bird et al. (2008) pointed out that OSS developer networks show a high degree of volatility, and cooperation structures tend to form *ad hoc* around current tasks in programming. Since then, many developments have emerged in the realm of open-source projects. For example, firms invest in external open-source communities, such as the Linux project or the GitHub platform, and private-collective innovation structures nowadays play an essential role in the open-source landscape (Homscheid et al., 2015; Liu et al., 2017). These developments may bring more

stability by strengthening the position of core programmers that are ready to invest considerable effort and can then make a living on open-source projects. Such developments may, however, also disturb established collaboration and communication structures when manipulating the focus of a project or sending new contributors. Whatever effect these developments may have on project success, they change at least the social and collaborative structures of a project (Homscheid et al., 2015).

Against the findings of Bird et al. (2008) one decade ago, recent studies showed that there are stable structures in the collaboration and communication habits of OSS projects. Joblin et al. (2017) investigated the dynamics of 18 large open-source projects and found that with time "the organizational structure of large projects is constrained to evolve towards a state that balances the costs and benefits of developer coordination" (Joblin et al., 2017, p.2050). The authors underlined the importance of three core principles of open-source software development that gain importance with the growing scales of a project: *scale freeness*, *modularity*, and *hierarchy*. *Scale-free networks* dispose of *hub nodes* with an extraordinarily large number of connections. This circumstance leads to the concentration of a large share of total coordination requirements to a small number of developers. This results in several beneficial characteristics including robustness and scalability (see also Dorogovtsev and Mendes, 2013). The local arrangement of nodes into groups that are internally well connected gives rise to a *modular structure*. Like scale-freeness, modularity helps to keep coordination needs among developers reasonable, and often arises naturally due to shared capabilities and interest in the same problem or challenge. The global arrangement of nodes into a layered structure, where small cohesive groups are embedded within larger and less cohesive groups, finally leads to the formation of *hierarchical structures* among core developers, while peripheral developers do not show such structures (Joblin et al., 2017).

So what are the influence factors that drive the formation of stable cooperation structures in open-source software projects? Eisenberger et al. (2001) found that one of the primary drivers of stable relationships between an individual and an organization is *reciprocity*, which is mainly driven by *perceived organizational support* (POS). There is a positive correlation between POS and an employee's loyalty, where loy-

alty is the felt obligation to support an organization's objectives and welfare. Shah (2006) found the responsiveness of an open-source project to be among the main motivation drivers in OSS projects – that is, the probability of keeping working augments with increased responsiveness. Observed and unobserved *homophily* leads to the formation of clusters, where homophily describes the phenomenon that two *similar* programmers are more likely to collaborate. The characteristics that determine similarity are either observed (a platform might decide to track whether programmers are male or female, or which infrastructure they use) or unobserved, like shared or complementary technical skills, and the interest in similar algorithms and problems. Stable modules and structures do not only arise due to shared interests: with time, strong social bonds within a limited group of developers arise, especially when the primary determinant of motivation is not material compensation (Shah, 2006).

Recent evidence on stability and social coherence in open-source projects has been provided, for example, by Ho and Rai (2017). The authors investigated the effect of monitoring measures on programmer participation and showed that quality assurance measures such as code acceptance and accreditation have a positive effect on a *volunteer's continued participation intentions*. The motivating effect of system feedback is crucial for both developers with short and long tenure, but feedback is fundamental for developers that only recently joined a program. There is a moderating effect of tenure on this phenomenon. The decision to continue contributing to an OSS project depends less on signals associated with the monitoring of contribution quality. In contrast, new members of the community are profoundly influenced in their decision by feedback mechanisms (e.g., comments on their role in the project). In this article, we seek to develop an adequate statistical model that can evaluate the four following central research questions:

RQ1: Are there organizational structures in OSS projects that help to predict future development of social stability in developer networks?

The developer network contains both information on the current activity level of programmers and collaboration habits. We seek to find out whether the modeling of

those dependency structures helps predict relationships between developers, relative to a setting where changes are assumed to be random.

RQ2: Do social ties and community structures in OSS projects become more stable over time or with the growing scope of a project?
As Joblin et al. (2017) argued, organizational principles and collaboration structures show predictable dynamic evolution, where differences arise across projects. This development assures a constant coordination overhead for single programmers and assures long-term viability. We validate whether there is evidence for the consolidation of the projects' organizational structure. We expect that relative predictive performance of methods that profit from a high degree of structural stability is higher compared to the performance of methods that imply little structural stability in later phases of the project.

2.2.2 Collaboration and Communication

Social bonds and clusters are likely to manifest not only in cooperation behavior but also in communication. Communication within open-source settings is "typically conducted in an open public manner and [...] stored for later reference" (Guzzi et al., 2013, p.277), hence easily accessible for research purposes. Our research complements previous work on how to improve development practices with regard to communication (e.g., Seaman, 1999). Bird et al. (2006, 2008) and Shihab et al. (2010) used metadata on e-mail communication (elements such as sender and receiver name, date, and time of communication) for the detection of social interaction structures in software development. They found that mailing list activity is related to coding activity. Consequently, there is a direct link between communication and collaboration tasks. Ogawa et al. (2007) used mailing list data to visualize the dynamic development of collaboration and communication structure of large, complex software projects. Guzzi et al. (2013) conducted a textual analysis of email list communication to analyze the interplay of collaboration and communication in OSS at a content level. The authors found that email lists are an essential, though not the only medium of programmer communication. They categorized the topics

of communication and found that only 16% of all threads were treating technical infrastructure. Social interaction (that is, threads about topics like social norms, contributors, acknowledgment, and coordination) is accountable for about 6% of all threads. Together with communication about project status (e.g., planning and communication about releases and due dates), the mentioned topics are accountable for around 13% of all email list communication. A significant share of communication does not directly imply implementation issues. This share can be assumed to reflect social relationships that go beyond the urgent technical needs of collaboration and programming (Guzzi et al., 2013).

Considering the importance of coordination in a virtual, distributed working environment, we seek to investigate the relationship between collaboration and communication, and ask two more research questions:

RQ3: Does the stability of social ties differ between collaboration and communication?

In both communication and coding, teams usually develop routines. An essential share of communication involves users of a software product, and in a considerable share of communication, no core developers are involved (Guzzi et al., 2013). We seek to compare the degree of stability in the communication domain to the degree of stability in the collaboration domain.

RQ4: Are there stable social bonds in communication behavior that go beyond *ad hoc* programming behavior?

Core developers are likely to build up an organizational "overhead", that reflects stable social ties that do not imply programming tasks (Guzzi et al., 2013; Joblin et al., 2017). Collaboration and communication are likely to be driven by similar social dynamics, and therefore to show a strong correlation. By accounting for one of the two as a predictor, we isolate structures in the data that go *beyond* the shared influences. For example, if the collaboration of programmers requires direct communication via email, then collaboration should be a good predictor of communication. By canceling out communication events that refer to collaboration,

and looking at the structures that persist, we isolate stable interaction habits in communication events that are not due to collaboration.

2.3 Statistical Modeling

To show whether there are organizational principles in the networks that influence the way people communicate and collaborate, we need to quantify the strength of social cohesion, as well as its stability over time. As the reasons or the nature of the modules are only implicitly observable in the network structure, we need an exploratory approach that does not require a priori parametric specification of the modular structures. Consequently, we quantify the modular structures in developer networks. We will next present an approach that explores social stability in an exploratory, flexible way. This approach, which allows for flexible parameter inference based on a Bayesian Monte Carlo method, is presented in Section 2.3.1. We then continue with an alternative dynamic model, which is less flexible and more mechanistic than the AME but allows for temporal extrapolation, prediction, and stability tracking. This approach is the *spectral growth model* (Section 2.3.2). As both models are inadequate to answer our research questions, we finally combine the two into a new model (Section 2.3.3).

2.3.1 Additive and Multiplicative Latent Factors

Social network analysis quickly became popular for investigating the organizational structures of software projects, for example, to find out about the determinants of software quality (Meneely and Williams, 2011). This is due to the fact that phenomena such as scale-freeness and modularity are linked to *relationships*, not individual characteristics of nodes, and therefore require techniques that can describe *dyadic*[1] data. Clusters, scale-freeness, modularity, and reciprocity describe relationships and cause cross-sectional dependency among developers. Statistical modeling has to take this dependency into account. For example, the activity level of a programmer i is informative about the probability that i will get into contact with another

[1] A *pair of nodes* is also called *dyad*, whereas *monadic* data describes *individuals* only.

programmer j – especially if i and j are part of the same highly connected cluster of programmers working on similar problems. This correlation between events violates the independence assumption for the single developers (Hoff et al., 2002).

In statistical terms, the phenomena described above (like modularity and reciprocity) are violations of independence between particular relationships and developers. Reciprocity, for example, means that the probability of observing two reciprocal ties is higher than the sum of the probability to observe a single tie: A person receiving an email is likely to respond to the person and message. Beyond the dependency between developers, there is also temporal dependency: Developers who have collaborated before are likely to collaborate again. Cross-sectional and temporal dependency violate assumptions of independence and exchangeability in relational data, which are the basis for the application of many statistical models (Hoover, 1982; Aldous, 1985). Statistical models such as Probit regression with a standard error structure assume independence, whereas network-related models take the dependency that arises between nodes explicitly into account (Hoff, 2009).

The *Additive and Multiplicative Effects* (AME) latent factor model (Hoff, 2008; Minhas et al., 2016a) accounts for cross-sectional dependency in dyadic data via integration of latent factor structures. Such exploratory interactive latent factor models have the core advantage of being able to handle dyadic cross-sectional dependency without explicitly modeling all tie formation mechanisms (Hoff, 2008). If these assumptions are correct, it is not necessary to explicitly find a measure for the reciprocal behavior of two nodes, but the estimation process will automatically infer the strength of this behavior from the observed data. This class of models is also handy when inference on the strength of the effect of a specific influence factor is of interest. Exploratory factorial models secure unbiased inference when parameters for observable covariates need to be estimated with dependent data. Accounting for unobserved latent factors prevents a bias in the estimation of slope parameters, even when the independence assumption is violated (Hoff, 2009).

The AME model describes tie formation processes in a network as being driven by observed and unobserved covariates and factors. Let $\mathbf{Y} = \mathbf{Y}(G)$ of dimension $N \times N$ be the unweighted adjacency matrix of a static graph or network G. That is, y_{ij}, the

relation between node i and j, can take on the values 1 or 0 depending on whether there is an *edge* (other notions for *edge* are *tie*, *arc*, or *link*) between two *nodes* (or *vertices*, in our case *programmers*) i and j, or not. y_{ii} is not defined, as the node cannot have a relationship with itself, and therefore filled with 0. Now let $P(\mathbf{Y})$ be the probability of observing a specific network's adjacency matrix $\mathbf{Y}$. Then, following Hoff (2009) and Fosdick and Hoff (2015), the probability distributions of the individual edges y_{ij} are conditionally independent of one another, given θ_{ij}, where θ_{ij} contains all observed and unobserved information about nodes i and j and their compatibility.

$$P(y_{ij}, y_{ik}, \ldots, y_{lk}|\theta_{ij}, \theta_{ik}, \ldots, \theta_{lk}) = P(y_{ij}|\theta_{ij}) \times P(y_{ik}|\theta_{ik}) \times \ldots \times P(y_{lk}|\theta_{lk}) \quad (2.1)$$

$$P(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{\alpha=1}^{n\cdot(n-1)} P(y_\alpha|\theta_\alpha) \quad (2.2)$$

Depending on how much information is available, different assumptions on the form of θ_{ij} can be made. First, we assume that θ_{ij} depends on some exogenous predictors $x_{1,ij}, x_{2,ij}, \ldots, x_{p,ij} = \mathbf{x}_{ij}$. But even if we account for these exogenous predictors, there may be some dependency left in the network, because not all information can be accounted for explicitly. Factor models can be used to mine the latent dependency structures for prediction and structural analysis purposes in an exploratory way, without the need to specify prior hypotheses on the reasons that drive cluster formation among developers (Hoff, 2009; Koren et al., 2009). θ_{ij} can be expressed as the sum of several effects driven by observed predictors and unobserved or *latent* factors:

$$\theta_{ij} = log\big(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\big) \quad (2.3)$$

$$\theta_{ij} = \boldsymbol{\beta}^T\mathbf{x}_{ij} + a_i + a_j + \mathbf{U}_{(i\cdot)}^T\mathbf{D}\mathbf{V}_{(j\cdot)} + \epsilon_{ij} \quad (2.4)$$

where $\mathbf{x}_{ij}$ are exogenous covariates, and can be specified to describe dyadic edge attributes, such as "both developers i and j have been on the project for more than

three years", as well as monadic covariates like "developer i has been contributing to the project for more than three years". If there are p observable covariates, $\mathbf{x}_{ij}$ is a vector of length p, just like $\boldsymbol{\beta}$. a_i and a_j denote additional random monadic effects, like the genuine interest of i and j in the project. Assume that both rows and columns of $\mathbf{Y}$ describe people, and y_{ij} describes a directed event such as "developer i sends an email to developer j". Then G is directed, $y_{ij} \neq y_{ji}$, and the random additive effect a_i and a_j are often interpreted as a measure of node i's "outgoingness" and node j's "popularity", respectively (Fosdick and Hoff, 2015). $\mathbf{U}_{(i\cdot)}$ denotes the i-th row of some matrix $\mathbf{U}$, and $\mathbf{V}_{(j\cdot)}$ denotes the j-th row of a matrix $\mathbf{V}$. The term $\mathbf{U}_{(i\cdot)}^T\mathbf{D}\mathbf{V}_{(j\cdot)}$ describes an unobserved multiplicative factor structure. $\mathbf{U}_{(i\cdot)}$ and $\mathbf{V}_{(j\cdot)}$ are rows of $\mathbf{U}$ and $\mathbf{V}$, respectively. As sender nodes are organized in rows, and receivers in the columns, we also speak of $\mathbf{U}_{(i\cdot)}$ and $\mathbf{V}_{(j\cdot)}$ as row- and column-specific latent nodal attributes. While a column in $\mathbf{U}$ and $\mathbf{V}$ describes an unobserved *factor* or *reason* for tie formation, a row contains the *scores* of an individual node with regard to these factors. The diagonal matrix $\mathbf{D}$ contains the weights of the characteristics $\mathbf{U}_{(i\cdot)}$ and $\mathbf{V}_{(j\cdot)}$ in the complete network, that is it describes the importance of the particular patterns for the entire network (Hoff, 2009; Fosdick and Hoff, 2015; Minhas et al., 2016a).

The k-th factor is denoted by $\mathbf{U}_{(\cdot k)}\mathbf{D}_{(kk)}\mathbf{V}_{(\cdot k)}^T$ and therefore spans an $N \times N$ matrix, in which every pair of nodes has a value. As $\mathbf{D}$ is a diagonal matrix, the product of two non-assorted vectors $\mathbf{U}_{(\cdot l)}\mathbf{D}_{(lm)}\mathbf{V}_{(\cdot m)} = 0$. If two nodes i and j have a high correspondence in their factor structure, then θ_{ij} is large and the nodes i and j have a high probability to connect (Hoff, 2009; Koren et al., 2009; Minhas et al., 2016a). With the AME framework, Hoff (2009) provided a flexible approach to exploring the cross-sectional dependency structure in a network for static (that is structural) link prediction. They assume reasons for tie formation that cannot be accounted for via observable edge attributes $\boldsymbol{\beta}^T\mathbf{x}_{ij}$ to reveal themselves in interactive factor structures $\mathbf{U}_{(i\cdot)}^T\mathbf{D}\mathbf{V}_{(j\cdot)}$. The difference to other approaches that base on factorization for link prediction (often used for recommender systems, see Koren et al., 2009), is that the approach's benefit is to provide fair inference for slope parameters β. Both the inferential approach suggested by Hoff (2009) and more mechanistic, but also more

efficient factorization methods for prediction (compare Koren et al., 2009), account for unobserved patterns via the factor structure.

The unobservable part of the information on nodes i and j, $\mathbf{U}_{(i\cdot)}\mathbf{DV}_{(j\cdot)}$, can be inferred from the observed topological structure of the network: if a group of nodes is highly interconnected, we can infer that they have something in common without actually knowing what this "something" is. This information is then helpful for prediction. Therefore, research in Empirical Software Engineering can profit from the model for finding patterns in the relationship between developers, without the need to explicitly specify all mechanisms and reasons that lead to the formation of modules, or hierarchy. In the AME model, the modeling of dependency in the factor structure allows for unbiased estimation of $\boldsymbol{\beta}^T\mathbf{x}$ under the assumption that some sources of dependency cannot be modeled explicitly via an observed covariate $\mathbf{x}$. Moreover, despite its static nature, McGraw and Menzinger (2008) and Mitrović and Tadić (2009) pointed out that spectral analysis bridges the gap between topological and dynamic analysis of networks, as the topological state of a network is usually the result of a dynamic process. While describing the topology of a network, factorization provides insights into the generation process, by carving out the underlying patterns that have led to the observed structure. Latent factors, extracted for example via spectral decomposition, inherently represent clustering, local heterogeneity, and other characteristics such as maximum distance across the network, bottlenecks, and degree of randomness (Seary and Richards, 2003; Hoff, 2009). Factorization is also a central methodology for dimensionality reduction. Dimensionality reduction reduces noise in the data by mapping a N-dimensional network structure to a K-dimensional subspace (where $K < N$ and N is the number of nodes or programmers) (Kunegis et al., 2010). As it computes the similarity of nodes, it is also useful for graph visualization (Seary and Richards, 2003). However, the model discussed above does not reflect changes in the structure of networks, only the results of a presumably still present, unobserved, stable growth mechanism. As we want to quantify the changes in the dependency structure, as well as compare the changes in the importance of the patterns in time, we will next combine the model with a dynamic perspective.

2.3.2 The Spectral Growth Model

Static models like the one above cannot track changes in the groups and patterns that are represented by the factors. When clusters of developers dissolve, or new groups form that involve some developers that are already active in a different cluster, the AME model is unable to track this change in the underlying factorial structures. As it uses a single snapshot of the network structure, the factorization of the aggregated network structure results in a mixed-up combination of patterns within the same factor that does not reflect the actual mechanisms of tie formation. Temporal aggregation retains topological patterns while losing activity patterns and temporal correlations (Gauvin et al., 2014). The assumption of constant factorial structures "can benefit the convenience of model inference, but is irrational empirically" (Zhu et al., 2016, p.729).

Kunegis et al. (2010) departed from the example of graph kernels that describe simple growth mechanisms like "a friend of a friend (of a friend) is a friend" to explain the relationship between stable spectral growth and mathematical growth models. Graph kernels are functions that are directly applied to the adjacency matrix $\mathbf{Y}$. The functions are usually functions that easily profit from the algebraic characteristics of a spectral decomposition of $\mathbf{Y}$. The spectral theorem assures that it is possible to decompose any matrix via eigen- (for symmetric matrices) or singular-value decomposition (SVD). Then, when the graph kernel function $p(\cdot)$ is applied to the decomposed matrix, it simplifies to a linear function fitting problem. Assume for a moment that the chosen factor extraction method for the model above is a spectral decomposition of $\mathbf{Y}$, and there is no error structure in the data. Then, it is true that $\mathbf{Y} = \mathbf{UDV}^T$. Graph kernels profit from the mathematical rule that $p(\mathbf{Y}) = \mathbf{U}p(\mathbf{D})\mathbf{V}^T$, due to the orthonormality of $\mathbf{U}$ and $\mathbf{V}$[2].

When applied to a temporal prediction, the parameters of the graph kernel $p(\cdot)$ are optimized such that $\mathbf{Y}_{t-1}$ is mapped to $\mathbf{Y}_t$, where t is a time indicator with $t \in 1, \dots T$. Now under the assumption of constant latent patterns in $\mathbf{U}$ and $\mathbf{V}$, this mapping reduces to the mapping of $\mathbf{D}_{t-1}$ to $\mathbf{D}_t$. Assuming that the triangle closing

[2]Note that while Kunegis et al. (2010) used symmetric matrices for their argumentation, where $\mathbf{U} = \mathbf{V}$, they point out that their model also applies to directed and bipartite matrices.

is the dominant pattern in tie formation in both $t-1$ and t, then $\mathbf{U}_t = \mathbf{U}_{t-1}$ and $\mathbf{V}_t = \mathbf{V}_{t-1}$. This simplification also holds for other graph kernels or the combination of kernels: As long as the same functions drive growth in the same way, the factorial structures can be assumed to be constant. As the matrices $\mathbf{D}_{t-1}$ and $\mathbf{D}_t$ are diagonal matrices, finding a function $p(\mathbf{D}_{t-1})$ that maps $\mathbf{D}_{t-1}$ on $\mathbf{D}_t$ reduces to a simple linear function fitting problem, which scales very easily to large networks (Kunegis et al., 2010).

This way, Kunegis et al. (2010, 2013) combined methods based on matrix factorization (where they use a spectral decomposition of $\mathbf{Y}$ as factor extraction method) in a dynamic perspective by pointing out that, no matter which patterns drive tie formation, as long as these patterns are constant, the factorization of the respective adjacency matrix should result in the same latent factor values of nodes i and j, $\mathbf{U}_{(i\cdot)}$ and $\mathbf{V}_{(j\cdot)}$. Instead of merely relying on the assumption of spectral stability, the authors proposed a series of tools to check the stability of factorial structures over time. Suggesting that if the latent dimensions represent underlying patterns and clusters, the authors deduced that these patterns and clusters might gain or lose in importance independently from one another. Therefore, Kunegis et al. (2010) suggested using a simple extrapolation of factor growth for link prediction, allowing single latent dimensions of the network to grow at their own speed.

To answer our research questions, we profit from the method in two regards. First, the tools suggested by Kunegis et al. (2010) allow us to track the stability of the factorial structure (which tells us about changes over time in the dominant organizational structures). Second, we use the link prediction performance to judge the degree of stability. Therefore, we next explain the tools that track stability.

The inner product of $\mathbf{U}_{(i\cdot)}$ and $\mathbf{V}_{(j\cdot)}$ together with the respective weights of the K factor scores, $d_1, \ldots, d_K$, result in a score that is directly related (though, in the model of Hoff (2009) not linearly) to the likelihood of i and j to have a common node, that is $P(y_{ij} = 1)$. The outer product of $\mathbf{U}_{(\cdot k)}$ and $\mathbf{V}_{(\cdot k)}$ represents the k^{th} orthogonal dimension of $\mathbf{\Theta}$, where $\mathbf{\Theta}$ is the $N \times N$ matrix containing all values θ_{ij}. The growth of the importance of the pattern or dimension k results in higher weights of that pattern. It can therefore be translated into growth in $\mathbf{D}_{(kk)} = d_k$.

When growth from time $t-1$ to t is considered, this means that $d_{k,t-1} < d_{k,t}$: Over time, more ties appear in the network, that is its *intensity* grows, and with it the weights.

Suppose that (i) the random term ϵ_{ij} can be ignored and (ii) that $\mathbf{M} = \mathbf{UDV}^T$ is the only source of growth in the network. Moreover, assume that (iii) the structures themselves are stable, that is $\mathbf{U} = \mathbf{U}_{t-1} = \mathbf{U}_t = \mathbf{U}_{t+1}$ and $\mathbf{V} = \mathbf{V}_{t-1} = \mathbf{V}_t = \mathbf{V}_{t+1}$. Then, growth in latent factors can be translated into growth in $\mathbf{D}_t$, formally:

$$\mathbf{M}_t - \mathbf{M}_{t-1} = \mathbf{U}\mathbf{D}_t^*\mathbf{V}^T \text{ where} \tag{2.5}$$

$$\mathbf{D}_t^* = \mathbf{D}_t - \mathbf{D}_{t-1}. \tag{2.6}$$

Any graph kernel is based on this assumption, as any graph kernel assumes a constant growth pattern. Graph kernels impose a parametric function on the growth of the values $d_{1,t-1}, \ldots, d_{K,t-1}$ to $d_{1,t}, \ldots, d_{K,t}$ to predict $d_{1,t+1}, \ldots, d_{K,t+1}$. This implies that the k^{th} eigenvector at time $t-1$ is always mapped to the k^{th} eigenvector at time t. Kunegis et al. (2010) showed that instead of this restrictive assumption, a flexible extrapolation of the values to $t+1$ results in gains in predictive performance when growth is spectral but irregular. Irregular growth describes a situation when the underlying factorial structures (that is, the underlying alliances among programmers, growth drivers like topics and tasks, and other social constellations) in a network are stable, but individual structures grow or shrink independently. Kunegis et al. (2010) provide more examples beyond the "a friend of a friend is a friend" hypothesis, and further explanation of graph kernels.

The performance of the spectral growth model is a good indicator of the plausibility of the assumption of spectral growth, that is, the varying importance of single modules in a software product. The better the predictions work in comparison to approaches that do not rely on factorial stability, the more we are confident that structures are stable. Therefore, we next explain the implementation of predictions based on the model. Kunegis et al. (2010) always picked three consecutive adjacency matrices at a time, $\mathbf{Y}_{t-1} = \mathbf{Y}(G_{t-1}), \mathbf{Y}_t = \mathbf{Y}(G_t)$, and $\mathbf{Y}_{t+1} = \mathbf{Y}(G_{t+1})$.

First, we decompose $\mathbf{Y}_t$. The resulting factorial structure provides the basis for calculating the "stable" eigenvalues of $\mathbf{Y}_{t-1}$ with the cosine similarity of the k^{th} vector in time T and in time $T-1$. In general, the cosine similarity of two vectors is a measure of similarity of two vectors that describes the cosine of the angle between two vectors:

$$\frac{\mathbf{U}_{(\cdot k),t-1} \cdot \mathbf{U}_{(\cdot k),t}}{\|\mathbf{U}_{(\cdot k),t-1}\|_2 \|\mathbf{U}_{(\cdot k),t}\|_2} = \frac{\sum_{i=1}^{N} \mathbf{U}_{(ik),t-1} \cdot \mathbf{U}_{(ik),t}}{\sqrt{\sum_{i=1}^{N} (\mathbf{U}_{(ik),t-1})^2} \cdot \sqrt{\sum_{i=1}^{N} (\mathbf{U}_{(ik),t})^2}} \tag{2.7}$$

The orthonormality of the matrices $\mathbf{U}_{t-1}$, $\mathbf{U}_t$ (and $\mathbf{V}_{t-1}$ and $\mathbf{V}_t$ accordingly) means that $\|\mathbf{U}_{(\cdot k),t-1}\|_2 = 1$ and $\|\mathbf{U}_{(\cdot k),t}\|_2 = 1$, and therefore the formula for the computation of cosine similarity simplifies to the following form:

$$\mathbf{U}_{(\cdot k),t-1}^T \mathbf{U}_{(\cdot k),t}. \tag{2.8}$$

.

The "artificial" diagonalization of the matrices results in nondiagonal $\tilde{\mathbf{D}}$ matrices. The closer $\tilde{\mathbf{D}}$ is to a diagonal matrix, the better the assumption of regular spectral growth is fulfilled.

$$\tilde{\mathbf{D}}_{t-1} = \mathbf{U}_t^T \mathbf{M}_{t-1} \mathbf{V}_t = \mathbf{U}_t^T \mathbf{U}_{t-1} \mathbf{D}_{t-1} \mathbf{V}_{t-1}^T \mathbf{V}_t \tag{2.9}$$

If $d_{(j)t}$ is the j^{th} eigenvalue at time t, its estimated previous value at $t-1$ is

$$\hat{d}_{(j)t-1} = \left(\sum_i \mathbf{U}_{(\cdot i)t-1}^T \mathbf{U}_{(\cdot j)t} \right)^{-1} \sum_i \mathbf{U}_{(\cdot i)t-1}^T \mathbf{U}_{(\cdot j)t} d_{(i)t-1} \tag{2.10}$$

If growth was perfectly spectral (that is if the factors were constant over time), then $\tilde{\mathbf{D}}_t = \mathbf{D}_t$, as then the multiplication of $\mathbf{U}_{t-1}$ with the transpose of the $\mathbf{U}_t$ or vice versa results in a unit matrix $\mathbf{I}$. Therefore, in the case of spectral growth, $\tilde{\mathbf{D}}_{t-1}$

should ideally be close to a diagonal matrix, and its estimated value can be used to judge the plausibility of the spectral growth assumption.

For prediction of $\mathbf{Y}_{t+1}$, the growth or shrinkage of the weights $\tilde{\mathbf{D}}_{t-1}$ to $\tilde{\mathbf{D}}_t$ is extrapolated. As the technique does not impose a parametric function to this step, it gives every dimension the freedom to grow or shrink at its own speed. $\mathbf{Y}_{t+1}$ is finally needed to validate the predictive performance of the algorithm. Kunegis et al. (2010) linearly extrapolated the growth of the eigenvalues to predict the successive eigenvalue $\hat{d}_{(k)t+1}$. That allows every subcommunity, encoded by an eigenvector, to grow at their own speed.

$$\hat{d}_{k,t+1} = d_{k,t} + (d_{k,t} - \tilde{d}_{k,t-1}), \tag{2.11}$$

where $\hat{\mathbf{D}}_{t+1}$ is the diagonal matrix of all predicted eigenvalues $\hat{d}_{(k)t+1}$.

The implementation by Kunegis et al. (2010) only accounts for factorial structures without additional explanatory covariates. The spectral growth model is limited in flexibility by the usage of spectral decompositions for factor inference, a method that does not allow to differentiate between observable and unobservable tie formation drivers. When the entire matrix is to be decomposed without consideration of random effects or further exogenous covariates, like in Kunegis et al. (2010), spectral decompositions provide a solution that scales very well to large datasets (Koren et al., 2009).

Kunegis et al. (2010) covered a wide spectrum of use cases, explicitly including *collaboration networks* and *communication networks*. Moreover, the authors claimed their method is appropriate for social networks, authorship networks, rating networks, citation networks and folksonomies, with varying success regarding prediction enhancement. They also cover not only symmetric networks. The authors use networks with unweighted and weighted edges, negative edges (when the setting allows for enmity relationships), and both unipartite and bipartite networks. Depending on the field of application, the latent factors driving tie formation represent "communities", "patterns", "topics", "modes" or "preferences". All these interpretations

can be subsumed as unobserved "reasons" and "structures" of why and how nodes are linked to one another (Kunegis et al., 2010). High stability of $\mathbf{U}$ and $\mathbf{V}$ means high stability in all unobserved elements of OSS development structures.

2.3.3 Synthesis of Models

The *spectral growth model* proposed by Kunegis et al. (2010) assesses the stability of tie formation drivers over time. It predicts single latent factor dimensions to grow at their own speed, and to change in relative importance over time. In our study, this means that single modules of a software product involve a specific group of developers. The intensity in which people work on a single module increases and decreases eventually. It optimizes predictive performance (in comparison to more restrictive graph kernels) only when the driving mechanisms of tie formation themselves are very stable, and only when the relative importance of single latent dimensions varies over time (Kunegis et al., 2010). As we have shown above, the model does not take additional covariates into account and is inflexible in the application to alternative data operationalization, for example, if growth is not a realistic setting.
The method proposed by Kunegis et al. (2010) technically allows us to answer RQ1, RQ2, and RQ3, as we do not need to include covariates to answer those questions. However, for interpretation, we benefit from a separation of node characteristics and interactive effects, because the model suggested by Kunegis et al. (2010) does not allow us to separate a general activity level of a programmer from a stronger involvement in a cluster. In order to answer RQ4, we need to simultaneously model communication and collaboration, which requires the adoption of the inferential strategy suggested by Hoff (2009). The AME model allows us to include a predictor in the model – the model proposed by Kunegis et al. (2010) factorizes $\mathbf{Y}$ via spectral decomposition, and is thereby unable to account for the various kinds of dependency that the AME can differentiate, or to establish a connection between communication and collaboration. Therefore, we combine the *spectral growth* and the AME model, using the strategy of factor inference suggested in Hoff (2009) in combination with the dynamic tracing of factorial stability suggested in Kunegis et al. (2010). The hybrid model will be able to account for the structures in collaboration networks

as a predictor. Doing this leaves us only with the structures not explained by collaboration in communication. A comparison of the importance of those patterns for fit and predictive performance then allows us to judge their prevalence in the software developer networks.

Figure 2.1 summarizes our research framework, including references to the most central research on the topic of communication and collaboration among developers that we build upon. Let G_t and H_t represent the graphs of the email communication network and the collaboration network for time slice t, respectively, where $t \in 1, \ldots, T$. $\mathbf{Y}(G) \rightarrow \{\mathbf{Y}(G_1), \ldots, \mathbf{Y}(G_T)\}$ and $\mathbf{X}(H) \rightarrow \{\mathbf{X}(H_1), \ldots, \mathbf{X}(H_T)\}$ denote the adjacency matrices of the graph series. For the sake of simplicity, we denote $\mathbf{Y}(G_t) = \mathbf{Y}_t$ (communication) and $\mathbf{X}(H_t) = \mathbf{X}_t$ (collaboration). Ties in $\mathbf{X}_t$ result from the contribution of two developers to the same code feature. We track which programmer worked on which feature at what time. This way, we also track which programmers worked on the *same* feature and "collaborated". Network ties in $\mathbf{Y}_t$ represent responses of one programmer to an email thread that another programmer has either initiated or also responded to within time slice t. Both networks contain the full set of nodes or programmers; that is, there might be programmers that did not communicate in G_t or people that only communicated and did not collaborate with anyone in H_t. Section 2.4 provides more details.

We measure how similar the $\mathbf{U}_t$ and $\mathbf{V}_t$ are over time. To do so, we choose a model that estimates only one $\mathbf{U}$ and $\mathbf{V}$, and compare these factors to the time-dependent matrices. The step of estimating a shared factor structure, independently of whether there are exogenous predictors or not, therefore bases on the following basic model[3]:

$$\theta_{ijt} = a_{it} + a_{jt} + \beta x_{ijt} + \mathbf{U}^T_{(i\cdot)} \mathbf{D}_t \mathbf{V}_{(j\cdot)} + \epsilon_{ijt} \tag{2.12}$$

Table 2.1 provides an overview of our four core research questions and validation strategies. Concerning **RQ1**, we compare the performance of a predictive method that takes dependency structures into account with one that assumes that all changes

[3]Note that there is only one predictor x_{ijt} at a time (that is, communication or collaboration), and therefore $\boldsymbol{\beta}^T \mathbf{x_{ijt}}$ reduces to the product of two scalars, βx_{ijt}.

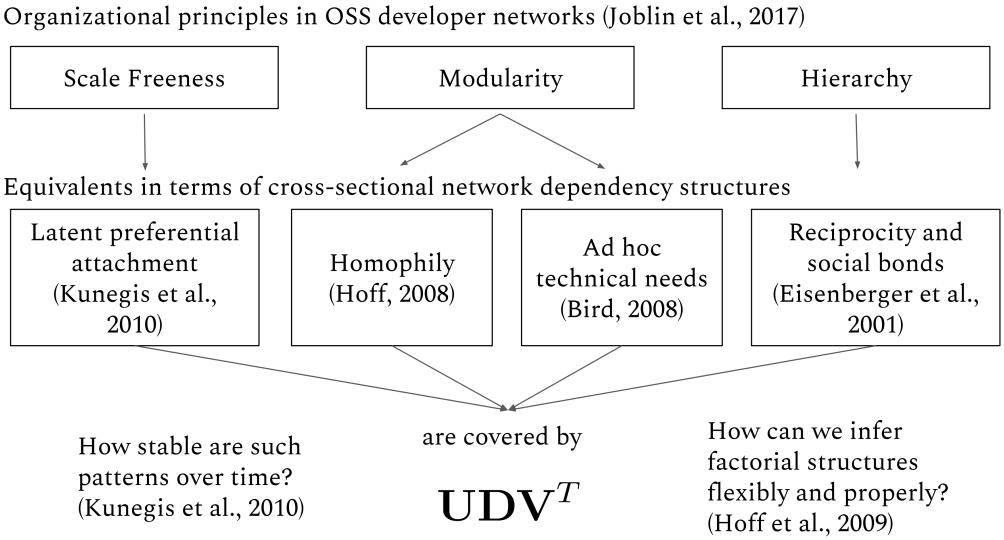

Figure 2.1: Representation of methodological context for case studies on factorial stability in OSS developer networks.

Table 2.1: Summary of research questions and validation strategies, based on the measurement of predictive performance and cosine similarity of core patterns.

	Benchmark	**Model**	**Evaluation strategy**
RQ1	$\hat{y}_{ij,t+1} = y_{ijt}$	$\theta_{ijt} = a_{it} + a_{jt} + \mathbf{U}^T_{(i\cdot)}\mathbf{D}_t\mathbf{V}_{(j\cdot)} + \epsilon_{ijt}$	Check performance with and without latent structures
RQ2	Performance at previous period	$\theta_{ijt} = a_{it} + a_{jt} + \mathbf{U}^T_{(i\cdot)}\mathbf{D}_t\mathbf{V}_{(j\cdot)} + \epsilon_{ijt}$	Evaluate performance over time and with growing scope
RQ3	$diag(\mathbf{I}_k) = \mathbf{1}$	$diag(\mathbf{U}^T_t\mathbf{U})$	Compare cosine similarity of latent patterns between communication and collaboration
RQ4	$\theta_{ijt} = a_i + a_j + \mathbf{U}^T_{(i\cdot)}\mathbf{DV}_{(j\cdot)} + \epsilon_{ijt}$	$\theta_{ijt} = \beta x_{ijt} + a_i + a_j + \mathbf{U}^T_{(i\cdot)}\mathbf{DV}_{(j\cdot)} + \epsilon_{ijt}$	Use collaboration as exogenous predictor, and compare weights $\mathbf{D}$ in latent factorial structure. Compare predictive performance when communication in $t+1$ is assumed to be known

are random, a *naïve* approach. This alternative approach assumes that the same events that occurred in time t will occur in time $t+1$ again, which essentially means that there is perfect stability, but no dynamic changes (growth or shrinkage). With regard to **RQ2**, we assess the stability of latent patterns with reference to project scale. We do that by judging the predictive performance of different approaches that imply different degrees of stability in growth. We further track the development of predictive performance with changes with growing scale of a project. We approximate project scale by the number of commits per three-month time window. Via the Bayesian implementation implied in the AME, we first estimate a shared $\mathbf{U}$ and $\mathbf{V}$ that spans the entire time line of the communication or collaboration networks:

$$\theta_{ijt} = a_{it} + a_{jt} + \mathbf{U}_{(i\cdot)}^T \mathbf{D}_t \mathbf{V}_{(j\cdot)} + \epsilon_{ijt} \tag{2.13}$$

$$\hat{\theta}_{ij,t+1} = \hat{a}_{i,t+1} + \hat{a}_{j,t+1} + \mathbf{U}_{(i\cdot)}^T \hat{\mathbf{D}}_{t+1} \mathbf{V}_{(j\cdot)}. \tag{2.14}$$

To prove that the model has value over a naïve approach, we compare the model's predictions to the simple heuristic $\hat{y}_{ij,t+1} = y_{ijt}$ (communication network) and $\hat{x}_{ij,t+1} = x_{ijt}$ (collaboration network). Then, the method for stability tracing described above is applied: We judge the stability of the patterns by applying different forecast methods whose success depends on the stability of the latent patterns. Predictions are calculated accordingly following the scheme:

$$\tilde{\mathbf{D}}_{t-1} = \mathbf{U}_t^T \mathbf{M}_{t-1} \mathbf{V}_t = \mathbf{U}_t^T \mathbf{U}_{t-1} \mathbf{D}_{t-1} \mathbf{V}_{t-1}^T \mathbf{V}_t \tag{2.15}$$

$$\hat{d}_{k,t+1} = d_{kt} + (d_{kt} - \tilde{d}_{k,t-1}) \tag{2.16}$$

$$\hat{a}_{i,t+1} = a_{it} + (a_{it} - a_{i,t-1}) \tag{2.17}$$

$$\hat{a}_{i,t+1} = a_{jt} + (a_{jt} - a_{j,t-1}) \tag{2.18}$$

Concerning **RQ3**, we seek to answer our question using tools for stability visualization suggested by Kunegis et al. (2010), a graphical representation of the matrices $\tilde{\mathbf{D}}_1, \ldots, \tilde{\mathbf{D}}_{T-1}$. The implementation described in Kunegis et al. (2010) departs from

a decomposition of the last matrix, which contains all the edges and nodes in the network, and then traces stability backward. Single patterns are more likely to be overlaid in this way. When using spectral decomposition, this means that we would have to add up all the time slices to receive an adjacency matrix that represents all nodes and edges. For the sake of comparison with the original method, we will also implement the spectral decomposition of a final, cumulated matrix, which corresponds more to the model suggested by Kunegis et al. (2010). However, as developer networks are highly dynamic, we do not wish to overlay the single matrices. With the inferential strategy proposed in Hoff (2009), it is possible to use the uncumulated matrices and estimate a shared factorial structure that bridges the single time slices without actually aggregating them to a single one. We compare the decomposition that results from this setting to one that uses the final, cumulated matrices that contain all the edges up to time T. To evaluate our setting, we base upon the following model, which estimates a single $\mathbf{UDV}^T$ despite a dynamic structure in the exogenous predictors (compare Hoff, 2017):

$$\theta_{ijt} = \beta x_{ijt} + a_i + a_j + \mathbf{U}_{(i\cdot)}^T \mathbf{D} \mathbf{V}_{(j\cdot)} + \epsilon_{ijt}. \tag{2.19}$$

Note the missing time index in $\mathbf{D}$. The advantage of using this setting is that it considers the time slices separately but estimates a single static interactive latent factor structure. $\mathbf{U}_{(i\cdot)}^T \mathbf{D} \mathbf{V}_{(j\cdot)}$ is static in this setting, any changes in the latent factor structure are reflected in the idiosyncratic error ϵ_{ijt}. The procedure counts repeated interaction events in every time slice separately, and results in an $\mathbf{UDV}^T$ that represents patterns that are present in all slices. We decompose the matrix according to equation (2.19) $\mathbf{UDV}^T$, which is then traced back even before $t-1$ over the entire time horizon of available data:

$$\tilde{\mathbf{D}}_t = \mathbf{U}^T \mathbf{M}_t \mathbf{V} = \mathbf{U}^T \mathbf{U}_t \mathbf{D}_t \mathbf{V}_t^T \mathbf{V} \tag{2.20}$$

The matrix $\mathbf{U}^T \mathbf{U}_t$ of dimension $K \times K$ contains on its main diagonal the simplified cosine similarity values as mentioned in equation (2.8). It is an indicator of how

similar the patterns in time t and over the entire series of networks are.

Concerning **RQ4**, we assess the gains in predictive performance that arise from using collaboration as a predictor for communication. Whether there are stable factor structures in communication behavior that go beyond joint programming means that there must be stable structures left in communication networks after accounting for joint programming. Consequently, we have to "get rid of" the factorial structures that are shared among collaboration and communication networks. To do this, we use x_{ijt} as predictor [4].

$$\theta_{ijt} = \beta x_{ijt} + a_i + a_j + \mathbf{U}_{(i\cdot)}^T \mathbf{DV}_{(j\cdot)} + \epsilon_{ijt} \tag{2.21}$$

where x_{ijt} and β are now scalars because there is only one predictor. Here, $\mathbf{X}_t$ plays the role of an independent predictor for $\mathbf{Y}_t$, in order to eliminate the shared factors from the data. To test the value of collaboration for prediction, we assume $x_{ij,t+1}$ to be unknown, and replace $x_{ij,t+1}$ in the prediction step by $x_{ij,t}$. To judge on how informative the *current* information and the usage of the factor structure are, we use a second setting, where we assume the exogenous predictor to be known for the forecast period $t+1$. With **RQ4**, we seek to make one part of the structure visible, by accounting for co-editing behavior in communication networks and approximating the "*ad hoc* technical need" patterns by co-editing behavior. When we use $\mathbf{X}_t$ as an explanatory variable, shared factorial structures cancel out, and only structures that are specific to $\mathbf{Y}_t$ remain.

Based on the latter approach, we will answer **RQ2** to **RQ4**. We structure the discussion in the following way:

- *Original algorithm*: Kunegis et al. (2010) discussed their prediction algorithm based on *growing* networks, where edges do not (usually) disappear again (that is, developers may join OSS projects, but may not leave). To demonstrate that we have implemented the algorithm by Kunegis et al. (2010), we will shortly discuss the results of an application of the matrix factorization on strictly

[4] Please note that we do not assume that the relationship between communication and collaboration is causal in any direction, but use the regression approach to cancel out the shared factor structures and the share of communication that is driven by collaboration needs.

growing networks, as originally suggested in Kunegis et al. (2010). There is no evaluation and comparison to other methods, as the implementation is only for demonstration of correct implementation, and better ease of understanding.

- To answer **RQ1**, we compare the prediction performance of an algorithm containing latent factorial structures to an algorithm not containing any interactive factorial structure. If the extrapolation of latent structures is beneficial for predictive performance, then we deduce that there are dyadic structures that are helpful for prediction.
- To answer **RQ2**, we track the predictive performance of a latent-factor-based prediction model over time and with a growing scope of a network. We approximate scope with the number of developers that are active within a time window t.
- To answer **RQ3**, we measure the cosine similarity of the latent factors. If there was more stability in collaboration than in communication, the stability of the factorial structures should be higher in the fist type of network.
- To answer **RQ4**, it is necessary to take on a multivariate perspective. Up to this point, we relied on the latent factorial structure to cover all the patterns that cause dependency in the networks. If both communication and collaboration are driven by certain technical needs, like the same programming tasks, then accounting for communication as a predictor should remove these structures. Therefore, we use communication between developer i and j at time t, x_{ijt} as a predictor, and evaluate the importance of the remaining factorial structure.

2.4 Empirical Study

To answer **RQ1** to **RQ4**, we will now present three empirical case studies concerning well-established open-source projects. Further details on data operationalization (Section 2.4.1) and model estimation and parameter inference (Section 2.4.2) are provided. Details on the validation of predictive performance (Section 2.4.3) provide the necessary insights into our evaluation criteria.

2.4.1 Subject Projects and Data Operationalization

We will now evaluate the proposed model of Section 2.3.3 in an empirical case study. We use version control system data on code contribution and email list communication of three open-source projects, namely QEMU[5], OpenSSL[6] and BusyBox[7] to empirically test our model. QEMU is particularly interesting, as it has imposed a special policy to their contributors according to which every patch must be sent to a mailing list first[8], which attributes a special role to the relationship between development and email list communication.

We use Codeface[9] (Joblin et al., 2015) to extract commit data from GitHub[10] repositories, and email data from the mailing list archive Gmane[11]. We denote the networks resulting from co-editing of source code as cochange networks, and email list networks as mail networks. Email communication is analyzed only among developers that have contributed to source code, that is, we neglect communication among users or between users and developers. Repeated collaboration or email contact within a three-month time window counts as a single collaboration, and we have a binary adjacency matrix for every time window. We use windows of three months length, and periods with fewer than ten communication events were excluded. When a period with more than ten events was followed by a period with less than ten events, the interim period was treated like a "normal" period to make sure that only consecutive time slices were compared. Table 2.2 provides an overview of the total number of active developers, co-edits on source code, email contacts, and first and last date analyzed. The limiting complexity factor was the number of mails, as there are relatively long periods at the beginning where no email communication was available. The minimum amount of communication rul affects a relatively long period, as can be seen in figure 2.A.1 in the Appendix.We implement our model and statistical reasoning in R (R Core Team, 2013), the Bayesian specification for AME is implemented

[5] www.qemu.org. QEMU is a virtual machine emulator.
[6] www.openssl.org. OpenSSL is an encryption library to secure connections on the Internet.
[7] www.busybox.net. BusyBox is a UNIX command-line tool suite.
[8] https://wiki.qemu.org/Contribute/SubmitAPatch, last access 11 December 2017
[9] https://siemens.github.io/codeface/
[10] https://github.com/, a project management platform basing on version control with Git
[11] http://gmane.org/, a mailing list archive

Table 2.2: Overview on the number of co-edits and email-based ties in the three subject projects, with start of the first and and end of the last three-month time slice.

Case study	Dev	Files	Co-edits	Mails	Start	End
`BusyBox`	221	1 138	12 945	3 353	2003-01-14	2016-01-14
`QEMU`	951	2 378	258 327	39 608	2003-04-29	2016-04-29
`OpenSSL`	164	1 161	8 330	980	2003-01-14	2015-10-14

in the `amen` package (Hoff et al., 2015).

Like Kunegis et al. (2010) and Joblin et al. (2017), we use a discretized time window approach, working on a stream of evolving developer networks derived from discrete software changes and mailing list communication. There are two series of networks for every case study, one for communication and one for collaboration. We use time slices of three months, indicating in a binary decision whether there has been email list communication or not between nodes i and j, and whether there was a collaboration between nodes i and j. The programmers (nodes) are related if they have edited a file together or responded to each other via email list. Figure 2.A.1 in the Appendix provides a first impression on the ratio of committing and communication within time slices of three months for the three case studies.

Within the empirical part of this article, we use only symmetric adjacency matrices (that is, $y_{ij} = y_{ji}$ and $x_{ij} = x_{ji}$). For symmetric $\mathbf{Y}$ and $\mathbf{X}$, the model reduces to a symmetric factorial structure $\mathbf{U}^T_{(i\cdot)}\mathbf{D}\mathbf{V}_{(j\cdot)} = \mathbf{U}^T_{(i\cdot)}\mathbf{D}\mathbf{U}_{(j\cdot)}$ (that is, $\mathbf{U} = \mathbf{V}$ and $\theta_{ij} = \theta_{ji}$). This means that when relationships are modeled symmetrically (programmers can only collaborate, there is no sender and receiver of help in programming or messages), the factorial structure simplifies. The simplification is straightforward, and for the sake of generality, as for future work, we want to consider asymmetric relationships, we keep using the asymmetric notation. First of all, we replicate the original algorithm proposed by Kunegis et al. (2010), which describes network *growth* in terms of the rising intensity of present nodes. Their model performs well in a setting where the intensity of network connections grows, and we would have to sum up the edges over time to create "artificial growth".

2.4.2 Parameter Inference

Hoff (2009) used spectral decomposition to illustrate the intuition of the multiplicative factor structure in the model while conducting the actual factor inference with an iterative Monte Carlo Markov Chain (MCMC) algorithm. There are two central reasons why Bayesian specification is better suited to our research context than spectral decomposition. First, there is no closed-form solution for the unbiased simultaneous estimation of a factor structure and the other parameters (like slope parameters for observed covariates) requiring an iterative algorithm for estimation (Bai, 2009), especially in the case when y_{ijt} is binary, and a generalized linear model is required. Second, spectral decompositions provide only point estimates, not providing any insights into the uncertainty structure of slope parameters and latent factors. Using an iterative Bayesian approach to factor inference allows us to make use of prior beliefs on parameter distributions, and to obtain estimators for uncertainty and confidence intervals.

Minhas et al. (2016a) provided a contextualization of package, model and method among other models for social network analysis. For Bayesian inference, the number of latent factors has to be fixed prior to the analysis, as the algorithms produce different estimates of $\mathbf{U}$ and $\mathbf{V}$ for different numbers of latent patterns (which is not the case for eigenvalue decomposition, where small eigenvalues can be sorted out later without changing the larger values). We checked the decrease of the absolute height of the eigenvalues, which usually resulted in four to five significant dominant latent components. The estimation of the unknown parameters $\boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{U}, \mathbf{V}$ and $\mathbf{D}$ bases on the specification of an adequate prior. Posterior distributions are obtained via Bayes rule,

$$P(\boldsymbol{\Theta}_t, \boldsymbol{\beta}, \mathbf{U}, \mathbf{D}_t, \mathbf{V} | \mathbf{Y}_t) \propto P(\mathbf{Y}_t | \boldsymbol{\Theta}_t, \boldsymbol{\beta}, \mathbf{U}, \mathbf{D}_t, \mathbf{V}) \times P(\boldsymbol{\Theta}_t, \boldsymbol{\beta}, \mathbf{U}, \mathbf{D}_t, \mathbf{V}). \quad (2.22)$$

Like spectral decomposition, the MCMC algorithm produces orthonormal columns in $\mathbf{U}$ and $\mathbf{V}$, that is, we scale the factors to a length of 1 and $\mathbf{D}_t$ keeps its interpretation as weights of the dimension spanned by pairs of columns across $\mathbf{U}$ and $\mathbf{V}$ (Hoff,

2009). An explicit treatment of Bayesian matrix factorization in a more general context is provided in Salakhutdinov and Mnih (2008). The iterative estimation for the given model starts with the specification of starting values $\boldsymbol{\psi}_0 = \{\boldsymbol{\Theta}_t, \boldsymbol{\beta}, \mathbf{U}, \mathbf{D}_t, \mathbf{V}\}$. Then, $\boldsymbol{\psi}$ is updated in several iterations to $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2$ etc. via MCMC. Therefore, $\boldsymbol{\beta}$ needs to be sampled from its conditional distribution $P(\boldsymbol{\beta}|\boldsymbol{\Theta}_t, \mathbf{U}, \mathbf{D}_t, \mathbf{V})$. Next, the k-th factors $\mathbf{u}_k, \mathbf{v}_k$ and their weights $\mathbf{d}_k$ are resampled from the respective conditional distribution (conditioning on everything but the k-th column or weight). Finally, $\boldsymbol{\Theta}_t^*$ is resampled, by using a random draw from a normal distribution for $\mathbf{E}_t$: $\boldsymbol{\Theta}_t^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{D}_t\mathbf{V}^T + \mathbf{E}_t^*$. Before starting a new iteration, θ_{ijt} is replaced by θ_{ijt}^*, with probability

$$\frac{P(y_{ijt}|\theta_{ijt}^*)}{P(y_{ijt}|\theta_{ijt})} \wedge 1, \tag{2.23}$$

where $\wedge$ denotes the upper ceiling of the probability value, that is, the probability cannot exceed 1 (100%). The generated samples of $\boldsymbol{\psi}_i$ converge with iterations to the posterior distribution for the observed information. Hoff (2009) provide an algorithmic representation of this procedure with some further methodological explanations. Bayesian inference bears the advantage that the posterior mean of the procedure is an average of all the parameters that are compatible with the observed data (Salakhutdinov and Mnih, 2008), thereby producing more general factor solutions than a spectral decomposition.

2.4.3 Cross Validation

A core challenge to our model setup is the fact that both matrix kernels and the original methodology used in Kunegis et al. (2010) base on *growing* networks for measuring predictive performance. Their implementation relies on a setting where networks are bound to grow in intensity, and edges "stay" once they have appeared. As the drop-out of programmers is a central element of stability in our networks, we do not wish to model data in this way. To be able to validate the correctness of our reproduction of the algorithm suggested by Kunegis et al. (2010), we split up our setting, and apply every prediction algorithm to the cumulated matrices as

well as to the original time slices. We will denote by y_{ijt} the ties in the original adjacency matrices, and by y'_{ijt} the ties in the cumulated settings (that is, $y'_{ijT} = 1$ if any $y_{ijt} = 1,\ t = 1, \ldots, T$). Accordingly, θ_{ijt} is the log odds of $y_{ijt} = 1$, and θ'_{ijt} is the log odds of $y'_{ijt} = 1$.

The baseline benchmark for predictive performance will be a *naïve* forecast. That is, we let the *spectral growth* model compete against a simple algorithm that simply predicts the same contacts that were there in the last period to also happen in the forecasting period. Table 2.A.1 provides an overview of all the models that will be estimated for validation of our research questions.

$$\hat{\mathbf{Y}}_{t+1,naive} = \mathbf{Y}_t \tag{2.24}$$

$$\hat{\mathbf{Y}}'_{t+1,naive} = \mathbf{Y}'_t \tag{2.25}$$

We denote the two algorithms `naive.sim` and `naive.cum` and compute them both for the communication and collaboration networks. Should this setting turn out to be a good predictor, then the information from $t-1$ is worthless, and there is no change in the latent patterns that is helpful for prediction. The hybrid model takes changes from $t-1$ to t into consideration. The growth of latent dimensions in that time is taken into account. The naïve model, however, only uses information from time t.

Our dataset is highly sparse and unbalanced, that is, there are many more ties "absent" than "present". Therefore, we use a measure of predictive accuracy that takes this fact into account. The Receiver Operating Characteristic (ROC) curve visually balances specificity (penalizes false predicted positives) and sensitivity (penalizes "missed" positives, that is, false negatives) of a prediction algorithm. An adequate single number summary of the shape of the ROC is the *Area Under Curve* (AUC), which can only take values between zero and one (Zhu, 2004). The AUC depends on both the True Positive Rate (TPR) and the False Positive Rate (FPR), basing on the predicted $\hat{\theta}_{ij,t+1}$. Many real-world network datasets are highly sparse, a fact that "has motivated the use of area under the ROC curve (AUC) as the de facto

performance measure for link prediction tasks" (Menon and Elkan, 2011, p.438), even though the AUC has also been criticized for ignoring the predicted probability values and the goodness-of-fit of the model (e.g., Lobo et al., 2008).

2.5 Results

Next, we will discuss the results of the algorithms proposed by Kunegis et al. (2010), to then present the results of our adaptations (that is, the adapted algorithm on time-window based networks with a Bayesian parameter inference). The evaluation of **RQ1** to **RQ4** with regard to predictive performance is summarized in table 2.A.2.

Replication of the Original Algorithm: The algorithm suggested by Kunegis et al. (2010) uses a spectral decomposition for factor inference. The original algorithm does not differentiate between the factorial structure, individual effects, and explicit influences and covariates. It moreover cannot deal with the drop-out of developers. Figure 2.2 shows the predicted behavior of the algorithm, similar to the results found by Kunegis et al. (2010): the weights of the single patterns/dimensions grow at individual speed, partially overtaking one another in importance over time. The black lines represent the $\tilde{\mathbf{d}}_{t1}, \ldots, \tilde{\mathbf{d}}_{t5}$, that is, the weights of the five most important latent dimensions at time t. Complementing figure 2.2, figures 2.A.5 and 2.A.4 show both settings `mail` and `cochange` for `QEMU` and `BusyBox`, respectively. The different shades of red in the second plot represent the order of the equivalent $d_{k,T}$. That is, the lines' colors encode the eigenvalue of the initial decomposition of $\mathbf{Y}_T$. The height of the red lines represents the cosine similarity of the eigenvectors. Especially the sudden surge in similarity at time $t = 16$ makes sense because, at the same time, the latent factors change in the order of importance (first panel of figure 2.2). The similarity of the eigenvector converges to one 1, which is a by-product of the computational method. As ever more edges are added to the network, and the importance of the single dimensions (represented in black) grows over time, overtaking one another occasionally. Figures 2.3 and 2.4 show the AUC performance of all models that were used to predict the cumulated `mail` networks. The plots show that the performance of the spectral growth model algorithm, as proposed by Kunegis

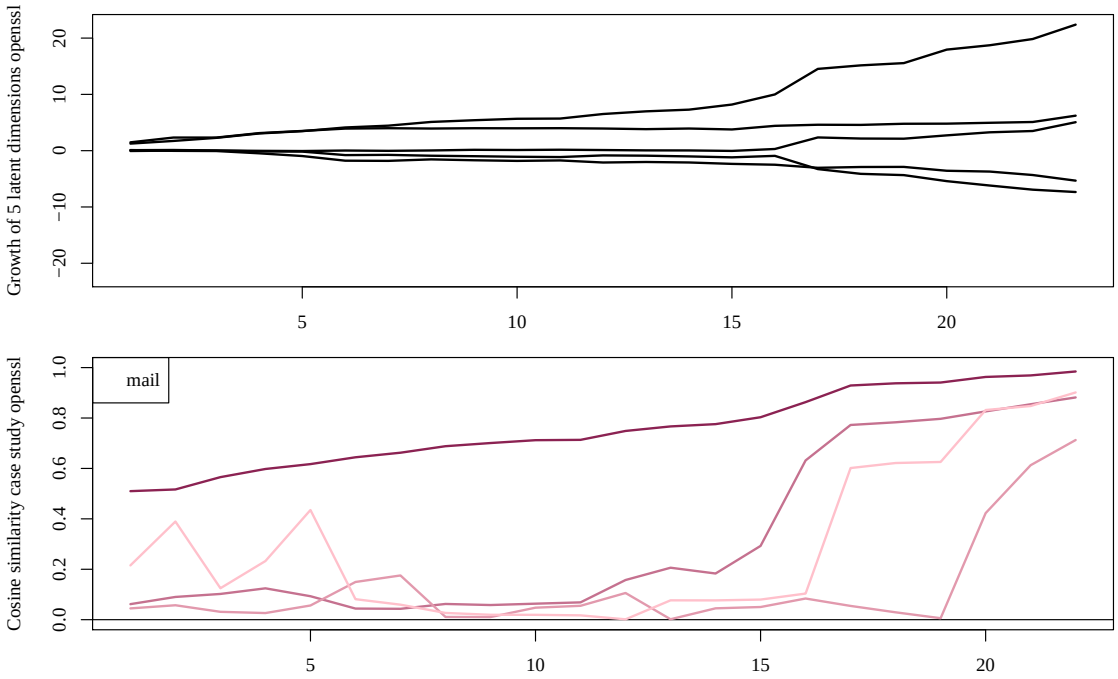

Figure 2.2: Growth of weights of the latent dimensions as resulting from spectral decomposition of cumulated matrices (panel 1), and cosine similarity of the first five dominant eigenvectors (panel 2) over time, for case study `OpenSSL` and the `mail` network.

et al. (2010) is pretty consistent across projects, and shows important differences across the `mail` and `cochange` networks. There is no significant improvement to the naïve forecast in the `cochange` networks, but important performance gains in the `mail` networks. The naïve prediction does not account for any form of growth. As the naïve model performs just as well as the model accounting for growth, the conclusion from our results is that there is no spectral growth (at least none that can be assessed via the growth of latent factors) in the `cochange` network. Additionally, the AUC is very close to 1, that is, the predictions from the naïve approach are quite hard to beat. In `mail` networks, there is growth that the latent factors account for. Figures 2.3 and 2.4 show that there are noteworthy trends in the growth of spectral structures in the `mail` network: The spectral growth model consistently outperforms the naïve model (the model that does not imply growth). For the collaboration-based network, the network shows a generally high level of predictability. The sudden drops in predictive performance in the middle of the project for the spectral decomposition-based approach suggest that the spectral prediction approach is less stable than the

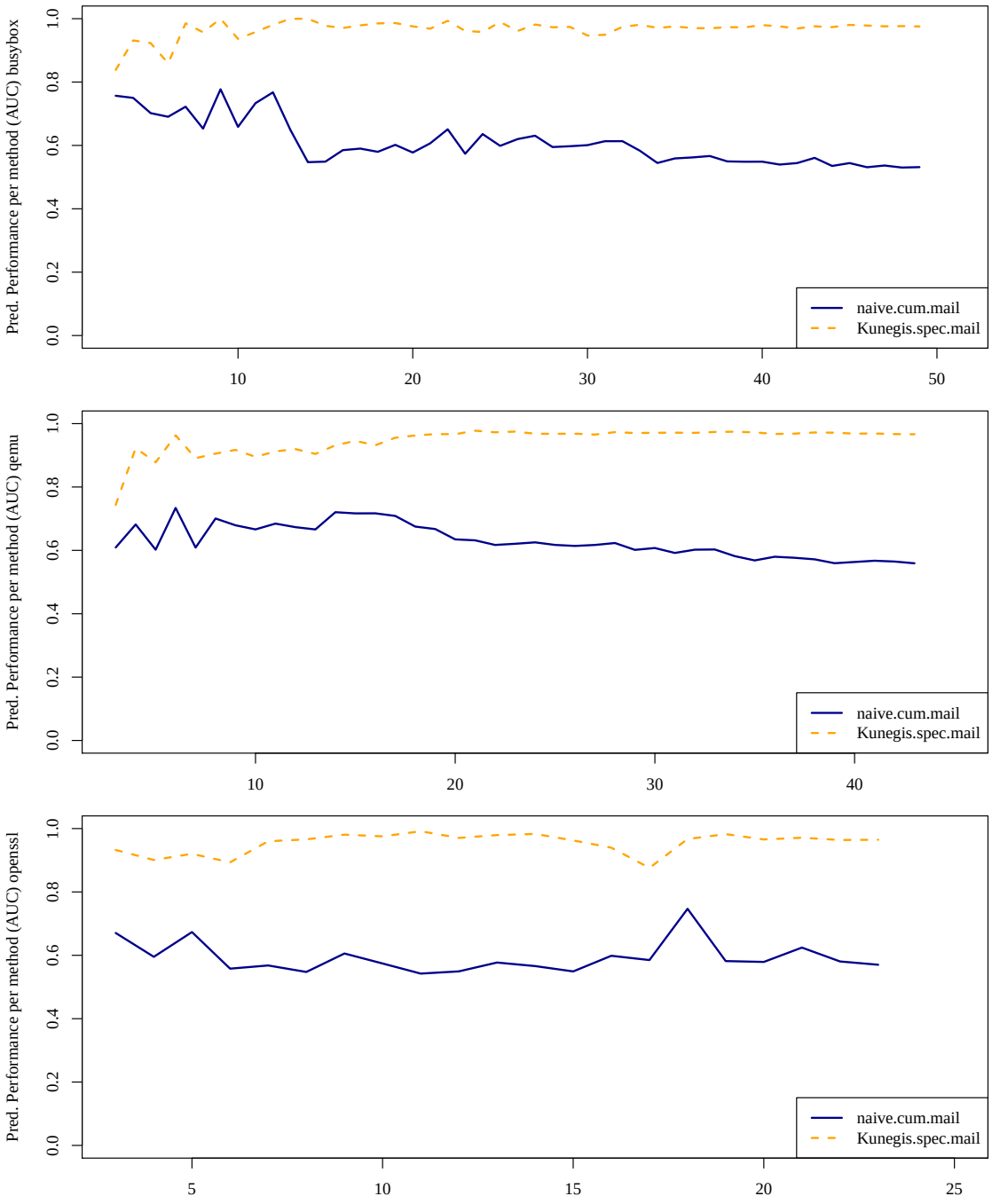

Figure 2.3: Predictive performance for analysis of cumulated `mail` networks, for the prediction of the cumulated adjacency matrices with information from the time t and $t-1$. Panel 1 refers to `BusyBox`, panel 2 to `QEMU` and panel 3 to `OpenSSL`.

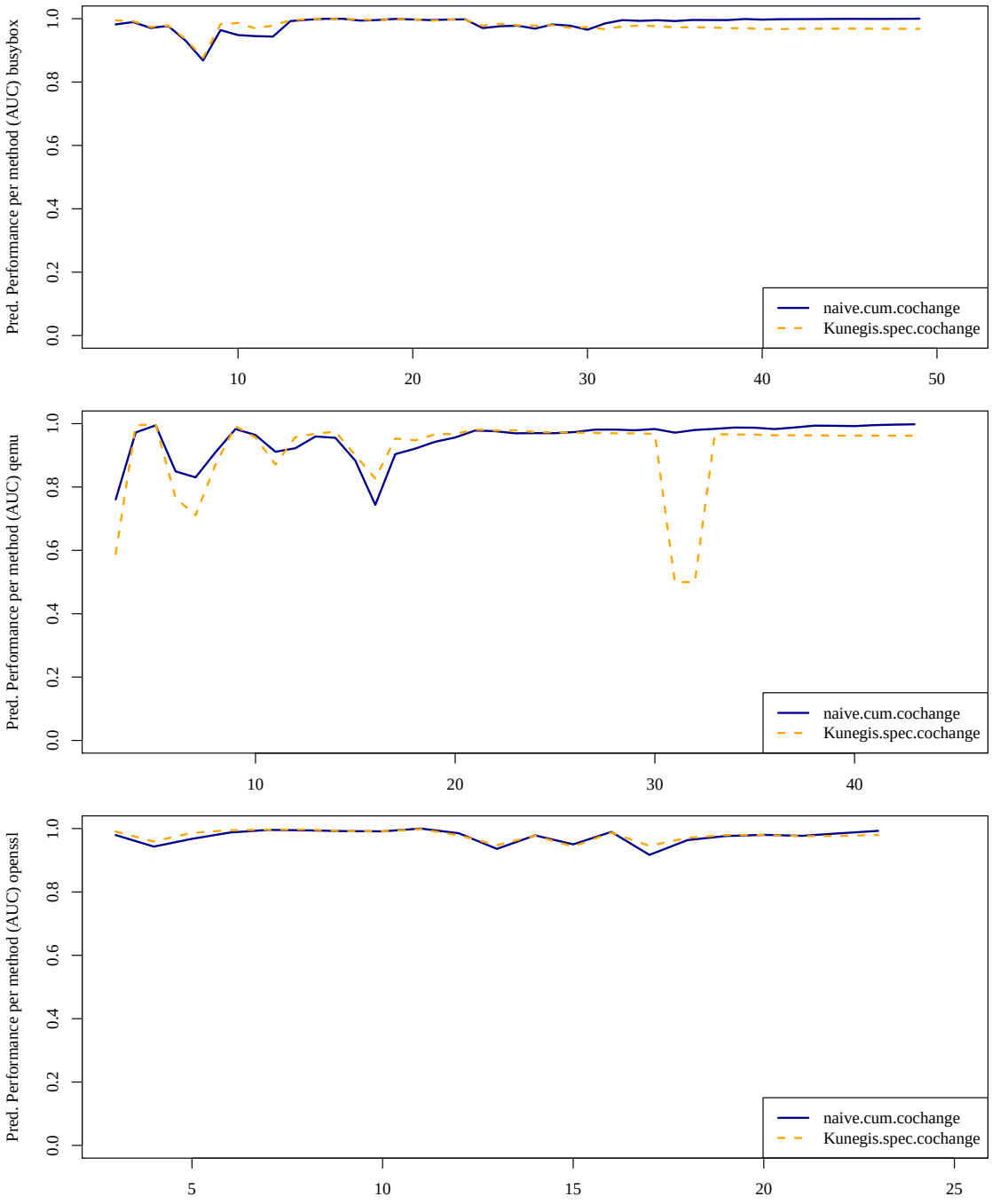

Figure 2.4: Predictive performance for analysis of cumulated `cochange` networks, for the prediction of the cumulated adjacency matrices with information from the time t and $t-1$. Panel 1 refers to `BusyBox`, panel 2 to `QEMU` and panel 3 to `OpenSSL`.

naïve approach. Table 2.A.2 summarizes the performance of the single approaches, indicating that with very high levels of AUC (0.9838 and 0.9765), both approaches describe the cochange network well (the naive approach outperforms the spectral approach and shows less variance, however). The mail network is badly described by the naive approach - with an AUC of 0.6040 for case study `BusyBox`) and a good AUC (0.9697 for case study `BusyBox`) for the spectral approach.

The **Replication** establishes a baseline for comparison, via application of the original method proposed by Kunegis et al. (2010). The setting where developers are not allowed to drop out from the project is highly predictable, and profits from the spectral growth assumption in that the performance of the prediction algorithm is much higher than the prediction performance of a naïve approach, which assumes $\mathbf{Y}_{t+1} = \mathbf{Y}_t$.

RQ1 - Presence of stable spectral structures Our first research question seeks to assess the importance of organizational structure in open-source developer networks. The higher the importance, the more engaging dyadic dependency structures via interactive latent factors should help predict the future state of a network in $t + 1$.

The three plots in figure 2.7 illustrate the predictive performance of the single models for predicting the non-cumulated adjacency matrices of the email networks. In the non-cumulated setting, the factorial models consistently outperform the naïve approach. The setting Kunegis.MCMC.mail represents the model suggested by Kunegis et al. (2010), enhanced by a Bayesian estimation strategy and additive effects a_i and a_j. Interestingly, the AUC measurements show that model also shows that using an MCMC-based inference with row effects and column effects does not help predict the models: The light blue dotted line performs worse than the spectral decomposition, which includes only the interactive factor structure, and no additional row effects. For the cochange, that is, the collaboration networks, the naïve method outperforms the model with row effects in some cases. However, the naïve method also shows the highest volatility. These results are confirmed in table 2.A.2.

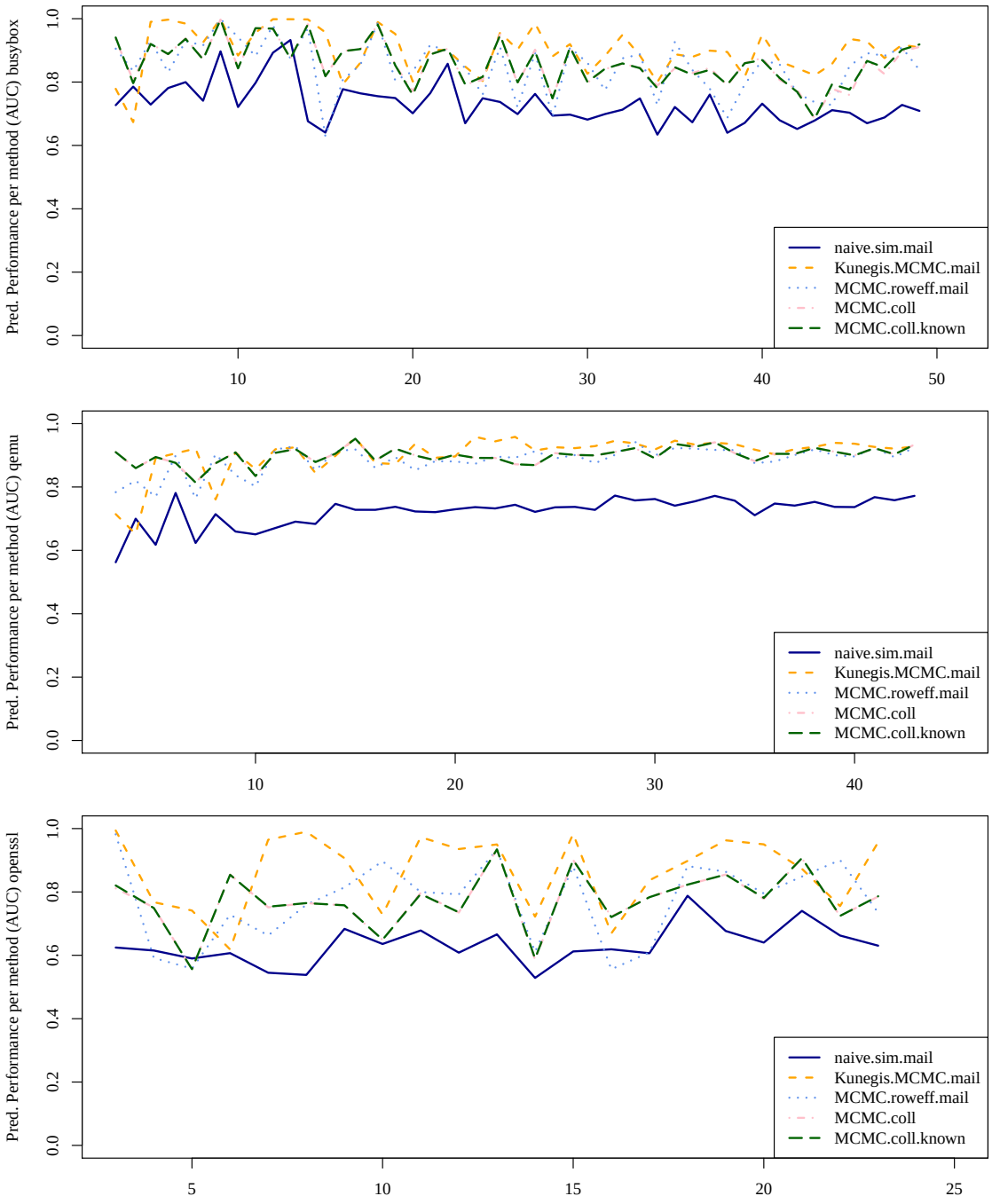

Figure 2.5: Predictive performance for analysis of uncumulated `mail` networks, for the prediction of the uncumulated adjacency matrices with information from the time t and $t-1$. Panel 1 refers to `BusyBox`, panel 2 to `QEMU` and panel 3 to `OpenSSL`.

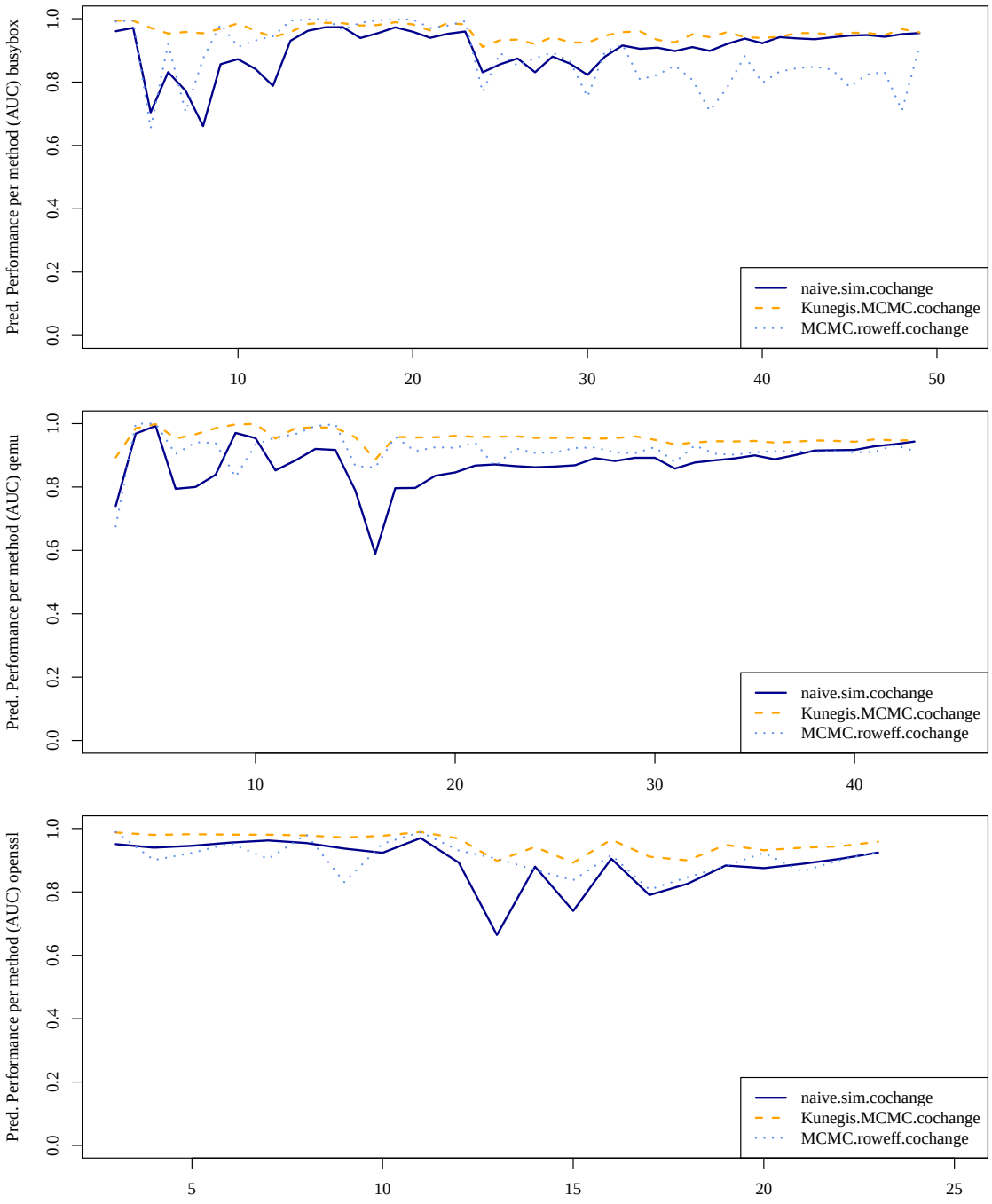

Figure 2.6: Predictive performance for analysis of uncumulated `cochange` networks, for the prediction of the uncumulated adjacency matrices with information from the time t and $t-1$. Panel 1 refers to `BusyBox`, panel 2 to `QEMU` and panel 3 to `OpenSSL`.

The naïve approach has an AUC of between 0.8763 and 0.9001 for the collaboration network, and 0.6336 to 0.7319 for the email network. Both for email and collaboration networks, the AUC values for the model involving interactive latent factor structures are consistently higher than the AUC values for the alternative methods.

Our results concerning **RQ1** suggest that interactive factorial structures, which represent the organizational structure, play an important role in link prediction. Their role is stronger for email than for collaboration networks. As the factorial approach is supposed to outperform the naïve approach only when there are dynamics in the growth of latent dimensions, our results suggest that the temporal dynamics from $t-1$ to t are more indicative about the network in $t+1$ in communication than in collaboration.

RQ2 – Growing stability over time or with growing scope With **RQ2**, we seek to figure out whether organizational structures in OSS projects become more stable – either over time or with the growing scope of a project. The answer is contained in a combination of figures 2.5, 2.6, figure 2.A.1 and table 2.A.2. The performance does not change significantly over time for case studies `BusyBox` and `OpenSSL`. For `QEMU`, there is a significant improvement in predictability over time, both for the naïve and for the alternative approaches. Figure 2.A.1 shows that for the periods where performance increases, the number of edges is also a lot higher. For higher number of edges, the factorial structures are more reliable to infer. It is therefore probably not the time, but the scope, that improves predictability.

There is growing predictability. However, this is likely to be related to the growing scope of the projects rather than to time itself. While only more case studies can help to differentiate both answers, we conclude concerning **RQ2** that spectral stability increases with time or with a growing number of developers.

To answer **RQ3**, that is, whether the stability of patterns differs between collaboration and communication, we check the cosine similarity of the dominant patterns. We cannot merely rely on predictive performance, as the predictability of both

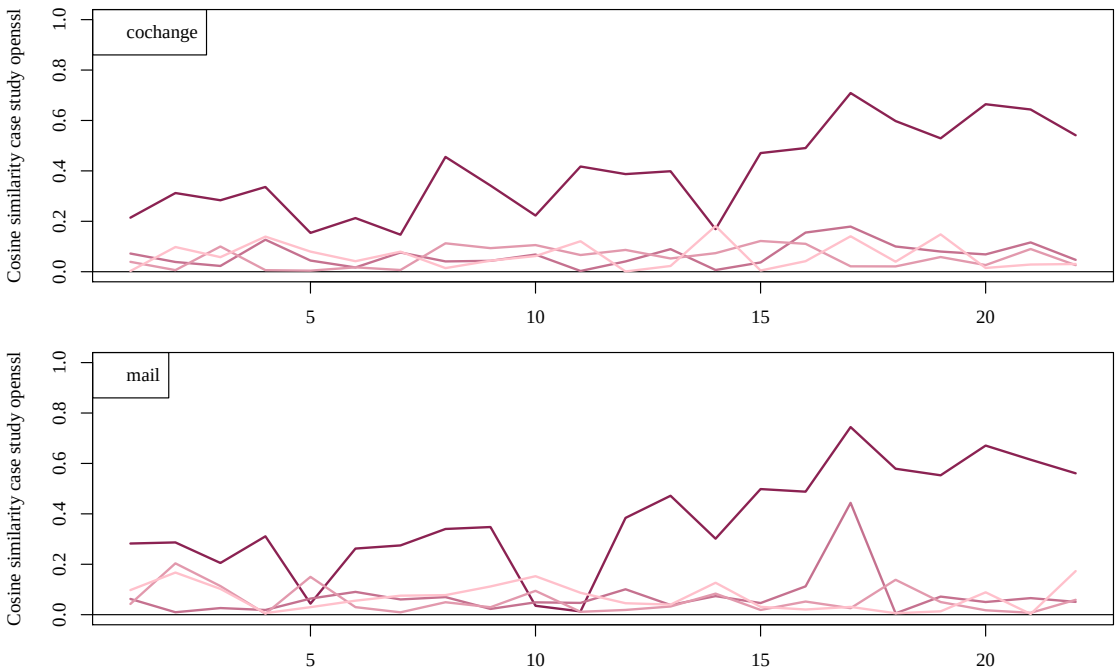

Figure 2.7: Comparison of cosine similarity of five core patterns, `cochange` (first panel) network versus `mail` network (second panel), for `OpenSSL`.

networks can be different. Moreover, the number of developers involved in both networks can influence the results. We consider the cosine similarity of the five most central patterns identified via the model with constant $\mathbf{UDV}^T$ according to equation (2.21), as represented in table 2.1.

If organizational structures were more stable in collaboration than in communication, then cosine similarity of factorial structures should be stronger in collaboration networks than in communication networks. Figure 2.7 shows a comparison of cosine similarity of the first five dominant patterns in the `cochange` networks (upper panel) and `mail` networks (lower panel) for case study `OpenSSL`. The same graphs for case studies `QEMU` and `BusyBox` can be found in figure 2.A.2 and figure 2.A.3, respectively. There are no significant differences in the stability of the core patterns between mail and cochange. There is one dominant pattern that is approximately stable from $t-1$ to t. As most of the variation in the adjacency is usually caused by the different out- and in-degrees, this result can be taken as a sign that the dynamics that showed to be beneficial for predictions because they reflect dynamics in the activity levels of the developers. It is interesting to see that the extrapolation of the growth of the

latent factors is helpful for prediction, but at the same time that the only stable dimension is the first one, which represents to a large extent the activity degree of the developers (which is usually the primary source of variance in a network). There is no difference in performance between the settings where the row effects are accounted for separately from the interactive factor structure, as can be seen in table 2.A.2.

Both the email and the collaboration networks show only one largely stable latent dimension. There are no significant differences in stability between the stability of communication and collaboration networks. The answer to **RQ3** is, therefore: No, there are no significant differences in stability between the two contexts.

RQ4 - Structures that go beyond ad hoc collaboration Finally, we want to verify whether there are social bonds in communication networks that go beyond ad hoc programming behavior (**RQ4**). For this purpose, we use a proxy for a part of the unobserved latent factor structure, which represents the *ad hoc* technical needs that drive communication. By accounting for this predictor, we eliminate the shared factor structure from the communication network: We do not assume that one is causal for the other. As table 2.A.2 and figure 2.5 show, knowing about the collaborative activities has little value for link prediction in the communication-based `mail` network. The AUC is augmented by as little as approximately 0.01 when the information is added to the model. More interestingly, it does not make a difference whether we use the information on collaboration in t (setting `MCMC.coll`) or in $t+1$ (setting `MCMC.coll.known`). Collaborative activity does not seem to have anything to do with who writes emails to whom, at least not when aggregated on a three-month window basis.

Adopting collaboration as an additional predictor for communication does not increase predictive performance, not even when we assume that collaboration in $t+1$ is known. The inclusion of contemporaneous information on communication does not improve the predictions of collaboration events. The answer to **RQ4** is negative; we do not find evidence for long-term stable social bonds in communication behavior that goes beyond collaboration.

2.6 Discussion and Outlook

To investigate the stability of the organizational structure of communication and collaboration networks in open-source software development, we have combined the static AME model (Hoff, 2009), and the dynamic *spectral growth* model (Kunegis et al., 2010). While the first provides maximum flexibility in estimation due to the Bayesian factorization, the second allows tracking the stability of the latent factorial structure. Static latent factor approaches like the one suggested by Hoff (2009) "automatically" model dependencies in networks. Often used as a black-box algorithm, their focus is on the optimization of structural, not dynamic, prediction. Kunegis et al. (2010) suggested combining a static matrix factorization with a flexible temporal extrapolation. This flexible growth of single "dimensions" of a network provides a new perspective on the possibility that the spectral description of network topology offers for the assessment of the dynamics of a network. However, the approach suffers from the impossibility of including other predictors in the decomposition: There is also a lack of "real-world" meaning of the latent dimensions. These drawbacks were partially compensated by our combination of the approach with a Bayesian estimation strategy and the inclusion of row effects and exogenous predictors. The next step to extend the model dynamically would be to determine the optimum time horizon over which stable structures are useful in predicting collaboration or communication behavior.

The fact that the spectral approaches outperform the naïve approach only in the `mail` networks and not in `cochange` networks, raises questions concerning the dif-

ference of the stability in both settings. The growth or shrinkage of particular dimensions is more indicative in `mail` settings, suggesting that the stability of latent patterns and their growth is more pronounced in this setting. While the naïve approach can take account of *constant* patterns, the spectral approach can mine the growth or shrinkage of latent patterns.

Our empirical results have shown that the added value for the prediction of new links in collaboration that can be gained from communication data is limited – suggesting that the purposes of email list communication are not directly linked to collaboration and programming tasks. To verify the general validity of these results, more case studies have to be considered, and other forms of communication, like communication via chat or GitHub issues, have to be analyzed as well. The primary goal of this research consisted in the modeling of stability of social ties in collaboration networks. Concerning the empirical case studies, a glaring flaw in our setting is the weak email activity in the first few years. We primarily use implementation-centered email lists, and most projects have several email lists for discussion of different topics. An extension of the data basis to other mailing lists would be reasonable for the discussion of social effects and reciprocity. The external validity of our findings can only be improved with more case studies and extension of the email list activity to fields where a social exchange is more likely. Most surprisingly, the results are remarkably consistent across case studies, indicating very similar levels of performance and variance in performance for all chosen approaches across case studies. The number of case studies should be extended to include some with fundamentally different behavior, to check the reaction of the method.

From a methodological perspective, there are three main potential fields of extension to our analyses. First, we can use directed networks for email communication. Second, an obvious choice for a model extension would be to include the lagged adjacency matrix (that is, $\mathbf{Y}_{t-1}$) as a predictor for contemporary communication, and to factorize the matrices under simultaneous consideration of both communication and collaboration, as suggested in Minhas et al. (2016b). Third, recently developed tensor-based methods have become feasible for ever-larger data settings. Communication and collaboration can,in such a context, be analyzed without as-

suming that one is the determinant of the other. This way, it is possible to assess how similar social structures are in both contexts (e.g., Anandkumar et al., 2014; Gauvin et al., 2014; Yuan et al., 2014; Hoff, 2015; Minhas et al., 2016b). The tensor-based perspective allows us to discard the one-directional influence of collaboration on communication and allows us to use information from several contexts (like collaboration and communication) simultaneously for parameter inference. Behavior in multiple domains can now be integrated (e.g., Fu et al., 2009; Boccaletti et al., 2014; Battiston et al., 2016). It seems, for example, promising to differentiate email lists regarding content and organize data as tensors. Such an approach would use a tensor decomposition for sparse data to distinguish several contexts of communication and analyze patterns for differences.

On a more abstract level, the analysis of collaboration and communication can be embedded in a more normative view on what structure *should* look like to achieve specific goals and to determine success factors for achieving such structures. The detection of structure benefits from the recent progress that allows analyzing several endogenous contexts of collaboration or communication at a time. Refining the temporal horizon of insight is then very important to differentiate different phases in projects, such as sprints and releases, and optimizing the phases in a project when high activity levels threaten code quality and coordination. To achieve this, approaches that do less rely on "black box" factorial inference are necessary, and that allow more normative, comparative perspectives on network structure.

2.7 Conclusions

The combination of a detailed static setting and a dynamic perspective on collaboration and communication networks allowed us to gain detailed insights into the stability of organizational structures of virtual OSS development working environments. Basing on the idea that stable structures manifest in constant latent factor structures, we combined the AME model suggested by Hoff (2009) with the spectral growth model proposed by Kunegis et al. (2010). We have validated the performance of prediction strategies that benefit from stable structures, as well as the

cosine similarity of underlying patterns over the single time slices. We validated our approach also by comparing it to the original algorithm for tracking spectral stability suggested by Kunegis et al. (2010).

Our results suggest that there are important dyadic structures in OSS project organization. The fact that algorithms that use information on current dynamics in latent dependency structures consistently outperform naïve approaches, which assume no growth in latent structures, shows that there is potential for knowing about group dynamics in programmer behavior. Moreover, row and column effects do not help predict collaboration and communication, meaning that it is not the current activity level of a programmer that determines his/her role in the network, but that it is the direct interplay of latent developer characteristics. Knowing about collaboration does not improve forecasts on communication, even when we assume that collaboration within the same time window is known – which might mean that there is more to communication than immediate coordination needs.

There is much need and potential for further investigation of the social and technical structures in OSS projects. Virtual communication and new forms of collaboration become ever more critical for contemporaneous work conditions. The exemplary character of OSS working conditions for other industries allows us to predict the probable development of the stability of social ties in similar organizational contexts. With many decades of data that OSS projects offer, and the public information policy, version control data hide actionable conclusions for the future development of our work structures. Who collaborated with whom, and to what extent collaboration is accompanied by stable communication patterns, helps in enhancing productivity and the mental sanity of digital workers. Research in this direction will help to improve communication and collaboration tools, providing, for example, good predictions on possible synergies via collaboration or recognizing stable relationships and knowledge synergies.

2.8 References

Aldous, D.J., 1985. Exchangeability and related topics, in: École d'Été de Probabilités de Saint-Flour XIII –1983. Springer, Berlin, Heidelberg, GER, 1–198.

Anandkumar, A., Ge, R., Hsu, D.J., Kakade, S.M., 2014. A tensor approach to learning mixed membership community models. Journal of Machine Learning Research 15, 2239–2312.

Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica 77, 1229–1279. doi:10.3982/ECTA6135.

Battiston, F., Iacovacci, J., Nicosia, V., Bianconi, G., Latora, V., 2016. Emergence of multiplex communities in collaboration networks. PLOS ONE 11, 1–15. doi:10.1371/journal.pone.0147451.

Bird, C., Gourley, A., Devanbu, P., Gertz, M., Swaminathan, A., 2006. Mining email social networks, in: Proceedings of the 2006 International Workshop on Mining Software Repositories, ACM, Shanghai, China. 137–143. doi:10.1145/1137983.1138016.

Bird, C., Pattison, D., D'Souza, R., Filkov, V., Devanbu, P., 2008. Latent social structure in open source projects, in: Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering, ACM, New York, USA. 24–35. doi:10.1145/1453101.1453107.

Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C.I., Gómez-Gardenes, J., Romance, M., Sendina-Nadal, I., Wang, Z., Zanin, M., 2014. The structure and dynamics of multilayer networks. Physics Reports 544, 1–122. doi:10.1016/j.physrep.2014.07.001.

Dorogovtsev, S.N., Mendes, J.F., 2013. Evolution of networks: From biological nets to the Internet and WWW. Oxford University Press, Oxford, UK.

Draheim, D., Pekacki, L., 2003. Process-centric analytical processing of version

control data, in: Proceedings of the Sixth International Workshop on Principles of Software Evolution, IEEE. 131–136. doi:10.1109/IWPSE.2003.1231220.

Eisenberger, R., Armeli, S., Rexwinkel, B., Lynch, P.D., Rhoades, L., 2001. Reciprocation of perceived organizational support. Journal of Applied Psychology 86, 42. doi:10.1037/0021-9010.86.1.42.

Fischer, M., Pinzger, M., Gall, H., 2003. Populating a release history database from version control and bug tracking systems, in: Proceedings of the International Conference on Software Maintenance 2003, IEEE. 23–32. doi:10.1109/ICSM.2003.1235403.

Fosdick, B.K., Hoff, P.D., 2015. Testing and modeling dependencies between a network and nodal attributes. Journal of the American Statistical Association 110, 1047–1056. doi:10.1080/01621459.2015.1008697.

Fu, W., Song, L., Xing, E.P., 2009. Dynamic mixed membership blockmodel for evolving networks, in: Proceedings of the 26th Annual International Conference on Machine Learning (ICML), 329–336. doi:10.1145/1553374.1553416.

Gauvin, L., Panisson, A., Cattuto, C., 2014. Detecting the community structure and activity patterns of temporal networks: A non-negative tensor factorization approach. PLOS ONE 9, 1–13. doi:10.1371/journal.pone.0086028.

Guzzi, A., Bacchelli, A., Lanza, M., Pinzger, M., van Deursen, A., 2013. Communication in open source software development mailing lists, in: Proceedings of the 10th Working Conference on Mining Software Repositories, San Francisco, USA. 277–286. doi:10.1109/MSR.2013.6624039.

Ho, S.Y., Rai, A., 2017. Continued voluntary participation intention in firm-participating open source software projects. Information Systems Research 28, 603–625. doi:10.1287/isre.2016.0687.

Hoff, P.D., 2008. Modeling homophily and stochastic equivalence in symmetric relational data, in: Platt, J.C., Koller, D., Singer, Y., Roweis, S.T. (Eds.), Advances in Neural Information Processing Systems. Volume 20, 657–664.

Hoff, P.D., 2009. Multiplicative latent factor models for description and prediction of social networks. Computational and Mathematical Organization Theory 15, 261–272. doi:10.1007/s10588-008-9040-4.

Hoff, P.D., 2015. Multilinear tensor regression for longitudinal relational data. The Annals of Applied Statistics 9, 1169–1193. doi:10.1214/15-AOAS839.

Hoff, P.D., 2017. Dyadic data analysis with amen. URL: http://cran.stat.upd.edu.ph/web/packages/amen/vignettes/amen.pdf.

Hoff, P.D., Fosdick, B., Volfovsky, A., He, Y., 2015. amen: Additive and Multiplicative Effects Models for Networks and Relational Data. URL: https://CRAN.R-project.org/package=amen. R package version 1.1.

Hoff, P.D., Raftery, A.E., Handcock, M.S., 2002. Latent space approaches to social network analysis. Journal of the American Statistical Association 97, 1090–1098. doi:10.1198/016214502388618906.

Homscheid, D., Kunegis, J., Schaarschmidt, M., 2015. Private-collective innovation and open source software: Longitudinal insights from Linux kernel development, in: Conference on e-Business, e-Services and e-Society, Springer. 299–313. doi:10.1007/978-3-319-25013-7_24.

Hoover, D.N., 1982. Row-column exchangeability and a generalized model for probability, in: Proceedings of the International Conference on Exchangeability in Probability and Statistics, North-Holland. 281–291.

Joblin, M., Apel, S., Mauerer, W., 2017. Evolutionary trends of developer coordination: a network approach. Empirical Software Engineering 22, 2050–2094. doi:10.1007/s10664-016-9478-9.

Joblin, M., Mauerer, W., Apel, S., Siegmund, J., Riehle, D., 2015. From developer networks to verified communities: a fine-grained approach, in: Proceedings of the 37th International Conference on Software Engineering (ICSE), IEEE Press. 563–573.

Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. Computer 42, 30–37. doi:10.1109/MC.2009.263.

Kunegis, J., Fay, D., Bauckhage, C., 2010. Network growth and the spectral evolution model, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), 739–748. doi:10.1145/1871437.1871533.

Kunegis, J., Fay, D., Bauckhage, C., 2013. Spectral evolution in dynamic networks. Knowledge and Information Systems 37, 1–36. doi:10.1007/s10115-012-0575-9.

Liu, M., Hull, C.E., Hung, Y.T.C., 2017. Starting open source collaborative innovation: the antecedents of network formation in community source. Information Systems Journal 27, 643–670. doi:10.1111/isj.12113.

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17, 145–151. doi:10.1111/j.1466-8238.2007.00358.x.

McGraw, P.N., Menzinger, M., 2008. Laplacian spectra as a diagnostic tool for network structure and dynamics. Physical Review E 77, 031102. doi:10.1103/PhysRevE.77.031102.

Meneely, A., Williams, L., 2011. Socio-technical developer networks: Should we trust our measurements?, in: Proceedings of the 33rd International Conference on Software Engineering, ACM, New York, NY, USA. 281–290. doi:10.1145/1985793.1985832.

Menon, A.K., Elkan, C., 2011. Link prediction via matrix factorization, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 437–452. doi:10.1007/978-3-642-23783-6_28.

Minhas, S., Hoff, P.D., Ward, M.D., 2016a. Inferential Approaches for Network Analyses: AMEN for Latent Factor Models. Technical Report. Cornell University. ArXiv preprint arXiv:1611.00460.

Minhas, S., Hoff, P.D., Ward, M.D., 2016b. A new approach to analyzing coevolving longitudinal networks in international relations. Journal of Peace Research 53, 491–505. doi:10.1177/0022343316630783.

Mitrović, M., Tadić, B., 2009. Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. Physical Review E 80, 026123. doi:10.1103/PhysRevE.80.026123.

Ogawa, M., Ma, K.L., Bird, C., Devanbu, P., Gourley, A., 2007. Visualizing social interaction in open source software projects, in: 6th International Asia-Pacific Symposium on Visualization (APVIS), IEEE. 25–32. doi:10.1109/apvis.2007.329305.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, AT. URL: http://www.R-project.org/.

Salakhutdinov, R., Mnih, A., 2008. Bayesian probabilistic matrix factorization using markov chain monte carlo, in: Proceedings of the 25th International Conference on Machine Learning (ICML), ACM. 880–887. doi:10.1145/1390156.1390267.

Seaman, C.B., 1999. Qualitative methods in empirical studies of software engineering. IEEE Transactions on Software Engineering 25, 557–572. doi:10.1109/32.799955.

Seary, A.J., Richards, W.D., 2003. Spectral methods for analyzing and visualizing networks: an introduction, in: Breiger, R., Carley, K., Pattison, P. (Eds.), Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers. National Academy of Science, 209–228.

Shah, S.K., 2006. Motivation, governance, and the viability of hybrid forms in open source software development. Management Science 52, 1000–1014. doi:10.1287/mnsc.1060.0553.

Shihab, E., Bettenburg, N., Adams, B., Hassan, A.E., 2010. On the central role of mailing lists in open source projects: An exploratory study, in: Nakakoji,

K., Murakami, Y., McCready, E. (Eds.), New Frontiers in Artificial Intelligence: JSAI-isAI 2009 Workshops, Springer Berlin Heidelberg, Tokyo, Japan. 91–103. doi:10.1007/978-3-642-14888-0_9.

Yuan, T., Cheng, J., Zhang, X., Qiu, S., Lu, H., 2014. Recommendation by mining multiple user behaviors with group sparsity, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 222–228.

Zhu, F., Chen, G., Heng, P.A., 2016. A Bayesian nonparametric approach to dynamic dyadic data prediction, in: 16th International Conference on Data Mining (ICDM), IEEE. 729–738. doi:10.1109/icdm.2016.0084.

Zhu, M., 2004. Recall, precision and average precision. Technical Report. University of Waterloo. Waterloo, Canada.

2.A Appendix

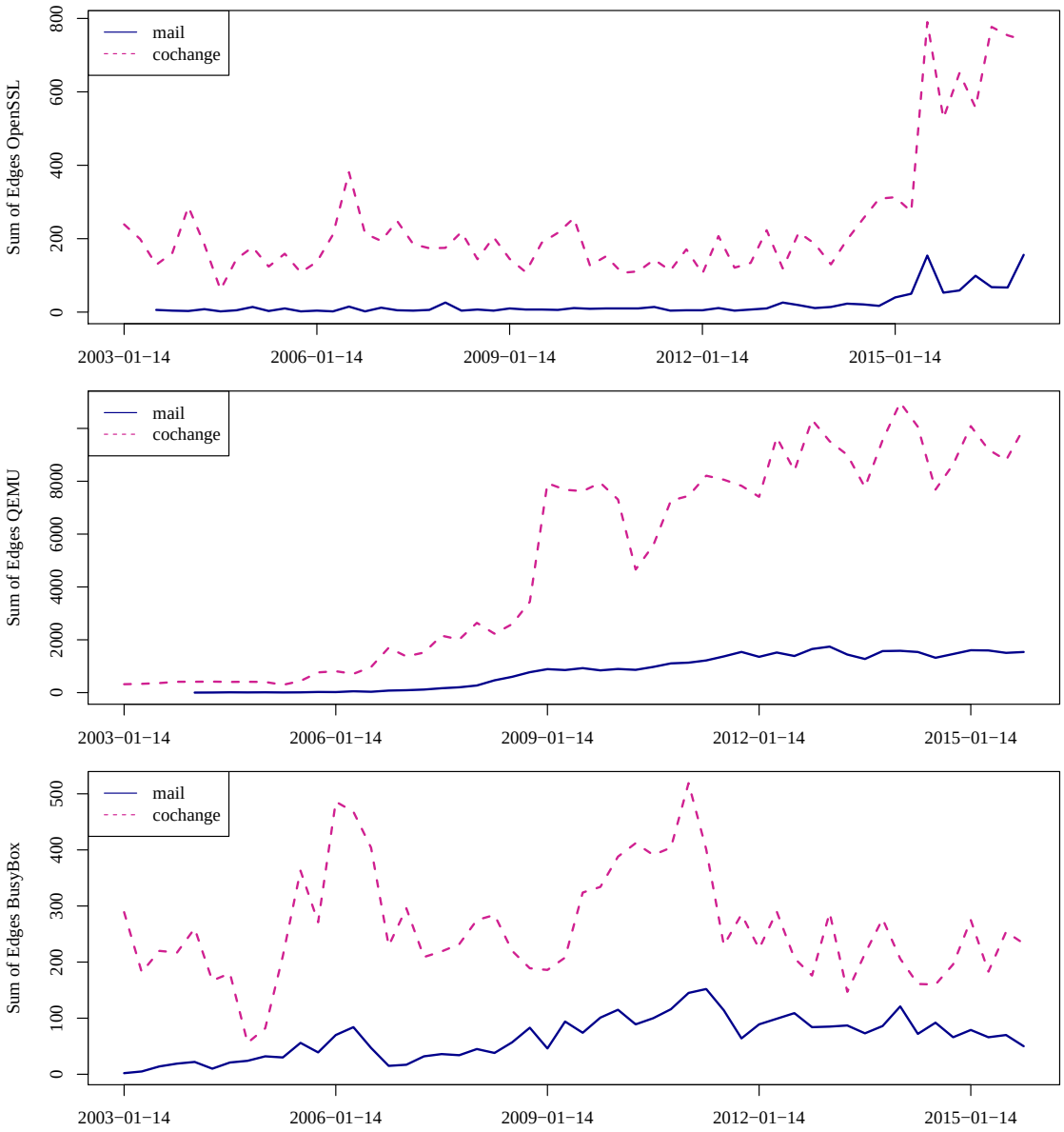

Figure 2.A.1: Comparison of email and source code editing activity, counted as total edges per time slice. Panel 1 refers to `OpenSSL`, panel 2 to `QEMU` and panel 3 to `BusyBox`.

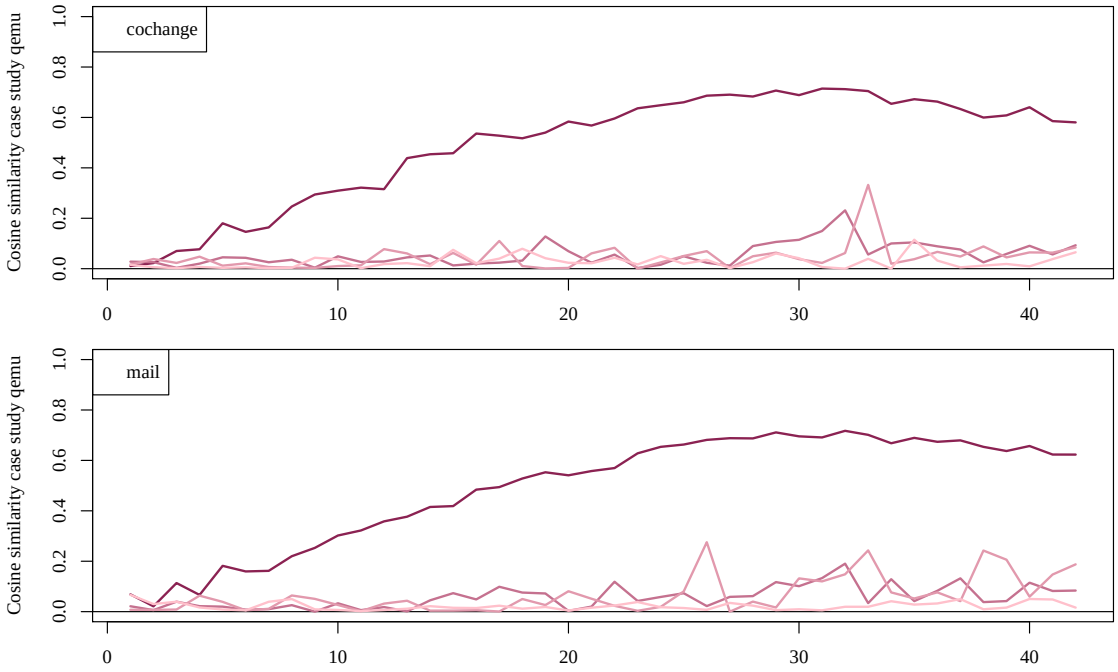

Figure 2.A.2: Comparison of cosine similarity of five core patterns, `cochange` network (first panel) versus `mail` network (second panel), for case study `QEMU`.

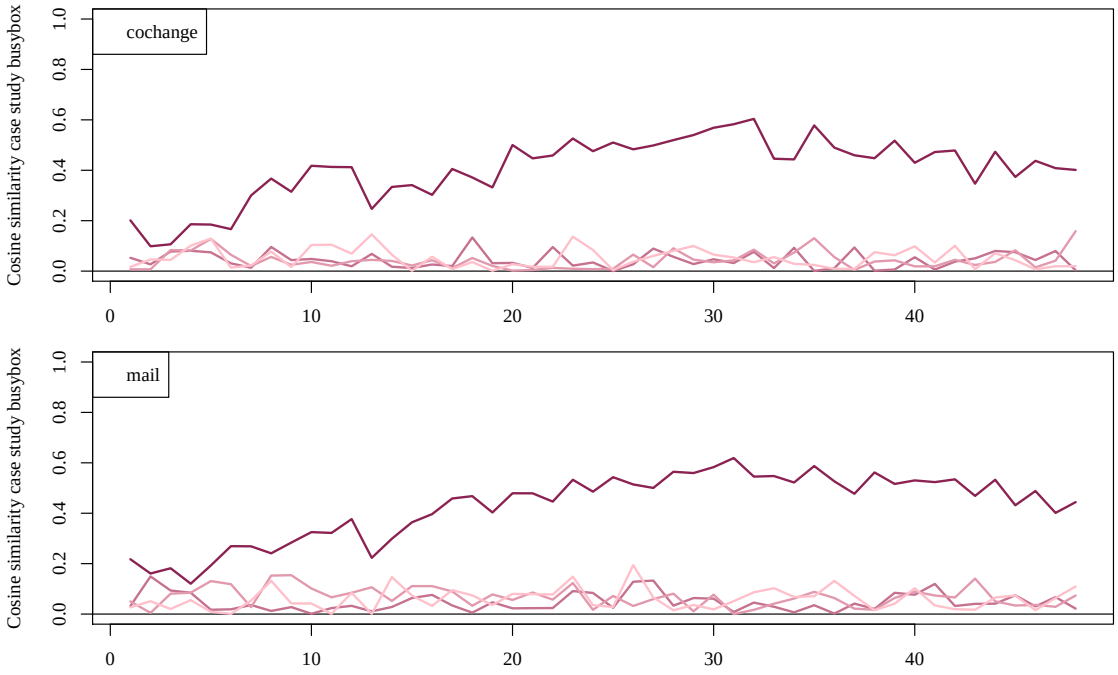

Figure 2.A.3: Comparison of cosine similarity of five core patterns, `cochange` network (first panel) versus `mail` network (second panel), for case study `BusyBox`.

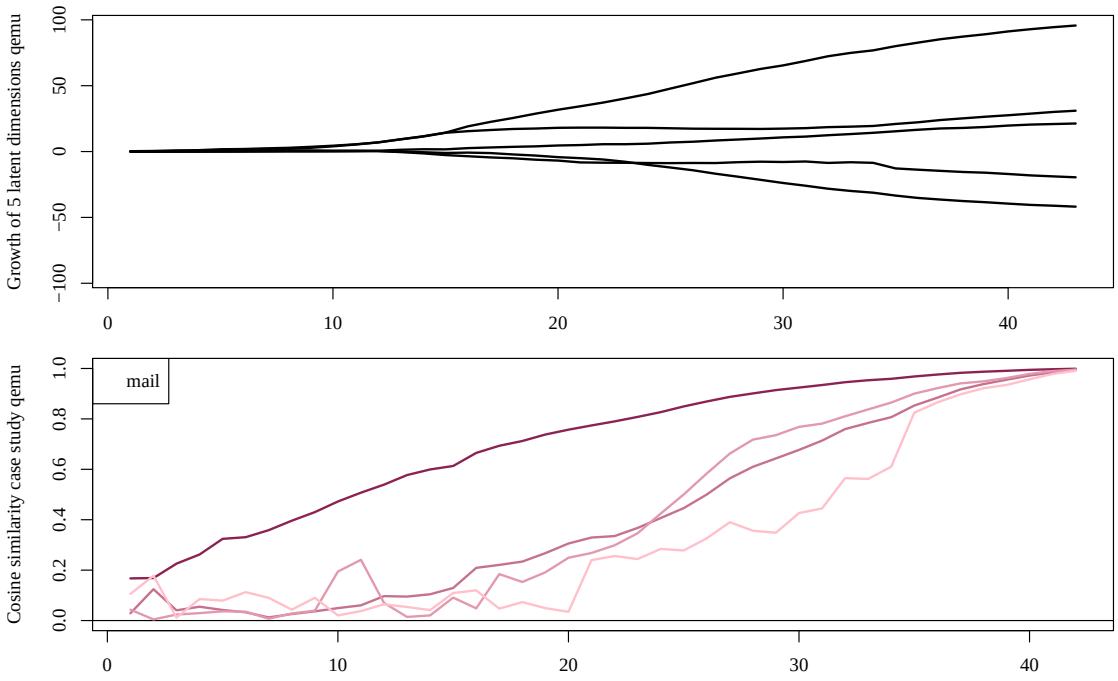

Figure 2.A.4: Growth of latent dimensions as resulting from spectral decomposition of cumulated matrices (panel 1 and 3), and cosine similarity of the first five dominant eigenvectors (panel 2 and 4) over time, for case study `QEMU`. The upper two panels represent the `mail` network, the lower two panels represent the `cochange` network.

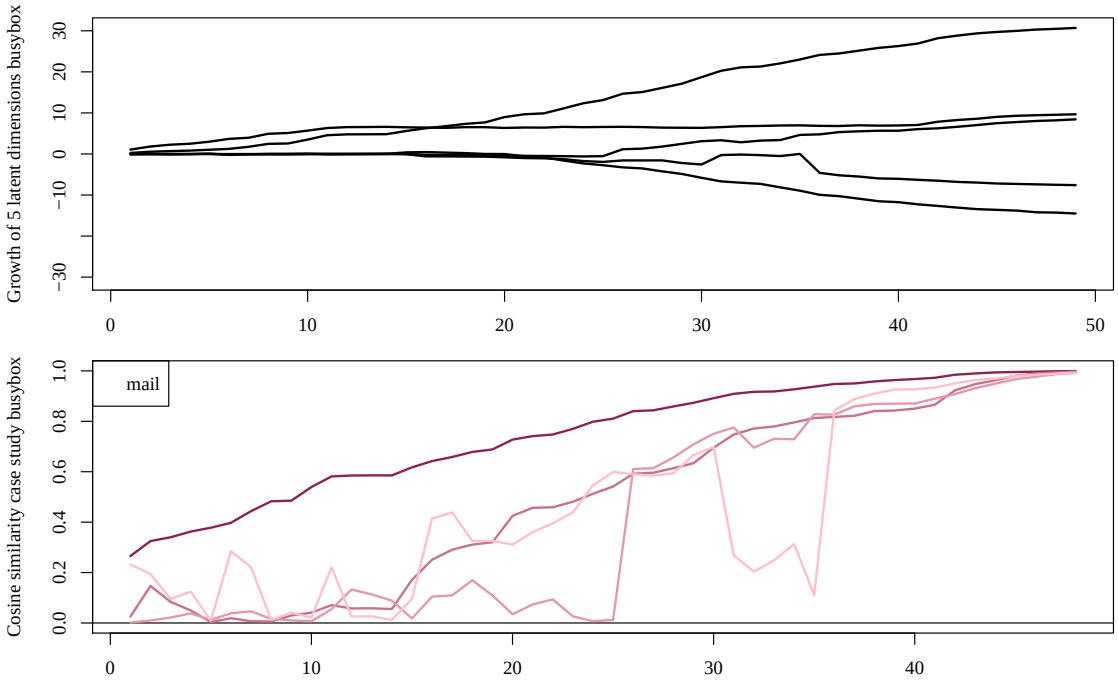

Figure 2.A.5: Growth of latent dimensions as resulting from spectral decomposition of cumulated matrices (panel 1 and 3), and cosine similarity of the first five dominant eigenvectors (panel 2 and 4) over time, for case study `BusyBox`. The upper two panels represent the `mail` network, the lower two panels represent the `cochange` network.

Table 2.A.1: List of the settings used for hypothesis validation. MCMC refers to a Monte Carlo Markov Chain based Bayesian implementation, $\mathbf{X}_{t+1}$ is an exogenous binary matrix referring to the collaboration matrix.

Setting name	Independent variable	Description/Estimation
`naive.cum`	$\mathbf{Y}'_t$ and $\mathbf{X}'_t$	Benchmark for prediction of cumulated matrices. Extensions `.mail` and `.cochange` indicate context.
`Kunegis.spec`	$\mathbf{Y}'_t$ and $\mathbf{X}'_t$	Prediction with the original method proposed by Kunegis et al. (2010), based on spectral decomposition. Extensions `.mail` and `.cochange` indicate context
`naive.sim`	$\mathbf{Y}_t$ and $\mathbf{X}_t$	Benchmark for prediction of simple matrices. Extensions `.mail` and `.cochange` indicate context.
`Kunegis.MCMC`	$\mathbf{Y}_t$ and $\mathbf{X}_t$	Parameter inference basing on eq. (2.21) without row effects, no exogenous predictors, based on MCMC. Extensions `.mail` and `.cochange` indicate context
`MCMC.roweff`	$\mathbf{Y}_t$ and $\mathbf{X}_t$	Parameter inference basing on eq. (2.21) with row effects, no exogenous predictors, based on MCMC. Extensions `.mail` and `.cochange` indicate context
`MCMC.coll`	$\mathbf{Y}_t$	Parameter inference basing on eq. (2.21) with row effects, $\mathbf{X}_t$ as predictor, based on MCMC.
`MCMC.coll.known`	$\mathbf{Y}_t$	Parameter inference basing on eq. (2.21) with row effects, $\mathbf{X}_t$ as predictor, assuming $\mathbf{X}_{t+1}$ to be known, based on MCMC.

Table 2.A.2: Results of a Cross-Validation on three month time windows of prediction of developer communication by case study, performance is indicated by mean and variance of Area Under Curve (AUC)

	Mean AUC busybox	Var. AUC busybox	Mean AUC qemu	Var. AUC qemu	Mean AUC openssl	Var. AUC openssl
`naive.cum.cochange`	0.9838	0.0006	0.9553	0.0028	0.9736	0.0006
`Kunegis.spec.cochange`	0.9765	0.0004	0.9238	0.0132	0.9550	0.0117
`naive.cum.mail`	0.6040	0.0045	0.6300	0.0027	0.5886	0.0023
`Kunegis.spec.mail`	0.9697	0.0005	0.9523	0.0008	0.9535	0.0013
`naive.sim.cochange`	0.8991	0.0048	0.8763	0.0046	0.8881	0.0062
`Kunegis.MCMC.cochange`	0.9562	0.0004	0.9576	0.0004	0.9519	0.0010
`MCMC.roweff.cochange`	0.8782	0.0086	0.9221	0.0013	0.9023	0.0024
`naive.sim.mail`	0.7319	0.0044	0.7270	0.0015	0.6336	0.0040
`Kunegis.MCMC.mail`	0.9017	0.0049	0.9085	0.0030	0.8523	0.0154
`MCMC.roweff.mail`	0.8481	0.0072	0.8868	0.0017	0.7168	0.0153
`MCMC.coll`	0.8581	0.0048	0.8993	0.0007	0.7910	0.0165
`MCMC.coll.known`	0.8584	0.0048	0.8992	0.0007	0.7909	0.0165

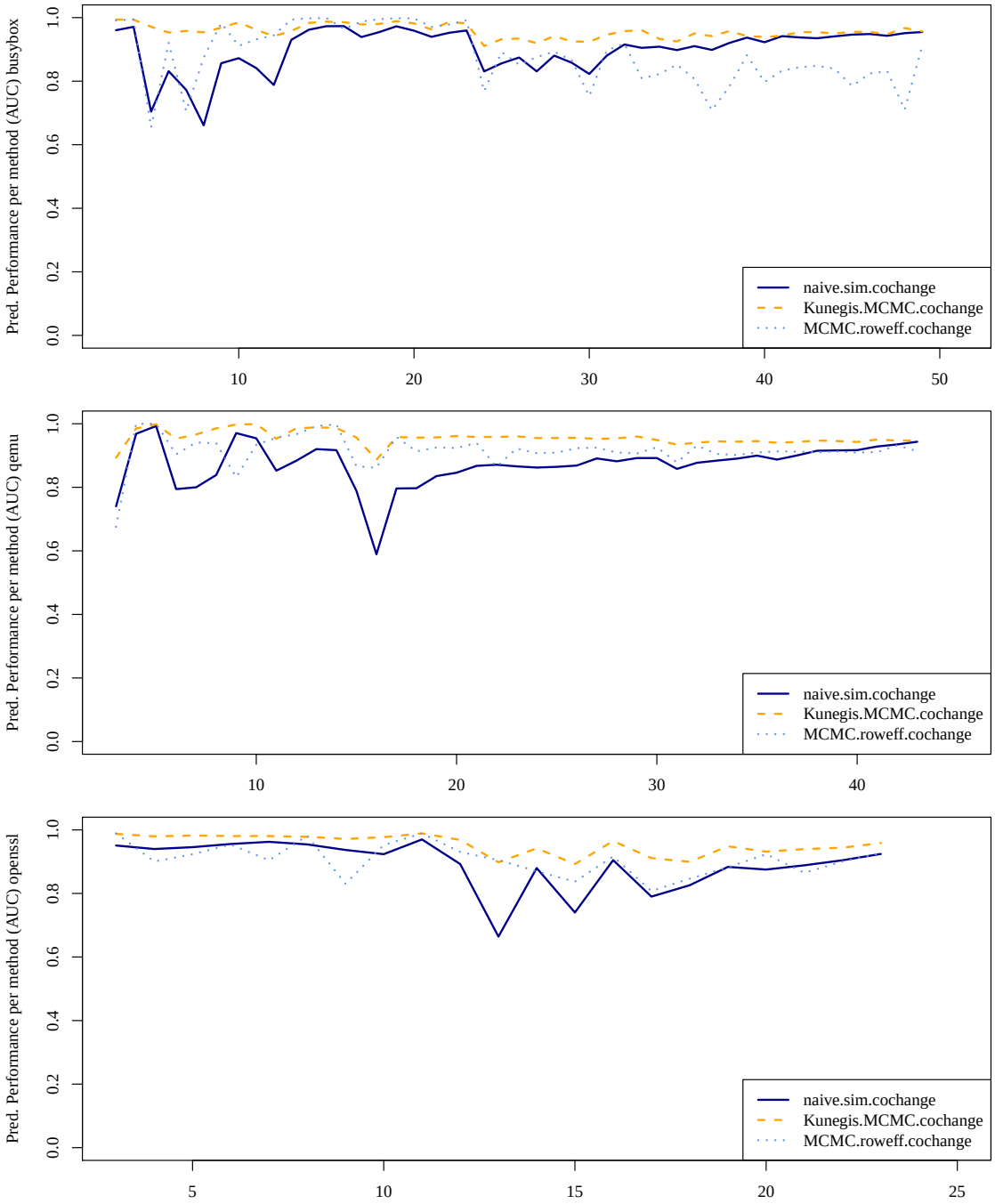

Figure 2.A.6: Predictive performance for analysis of uncumulated `cochange` networks, for the prediction of the uncumulated adjacency matrices with information from the time t and $t-1$. Panel 1 refers to `BusyBox`, panel 2 to `QEMU` and panel 3 to `OpenSSL`.

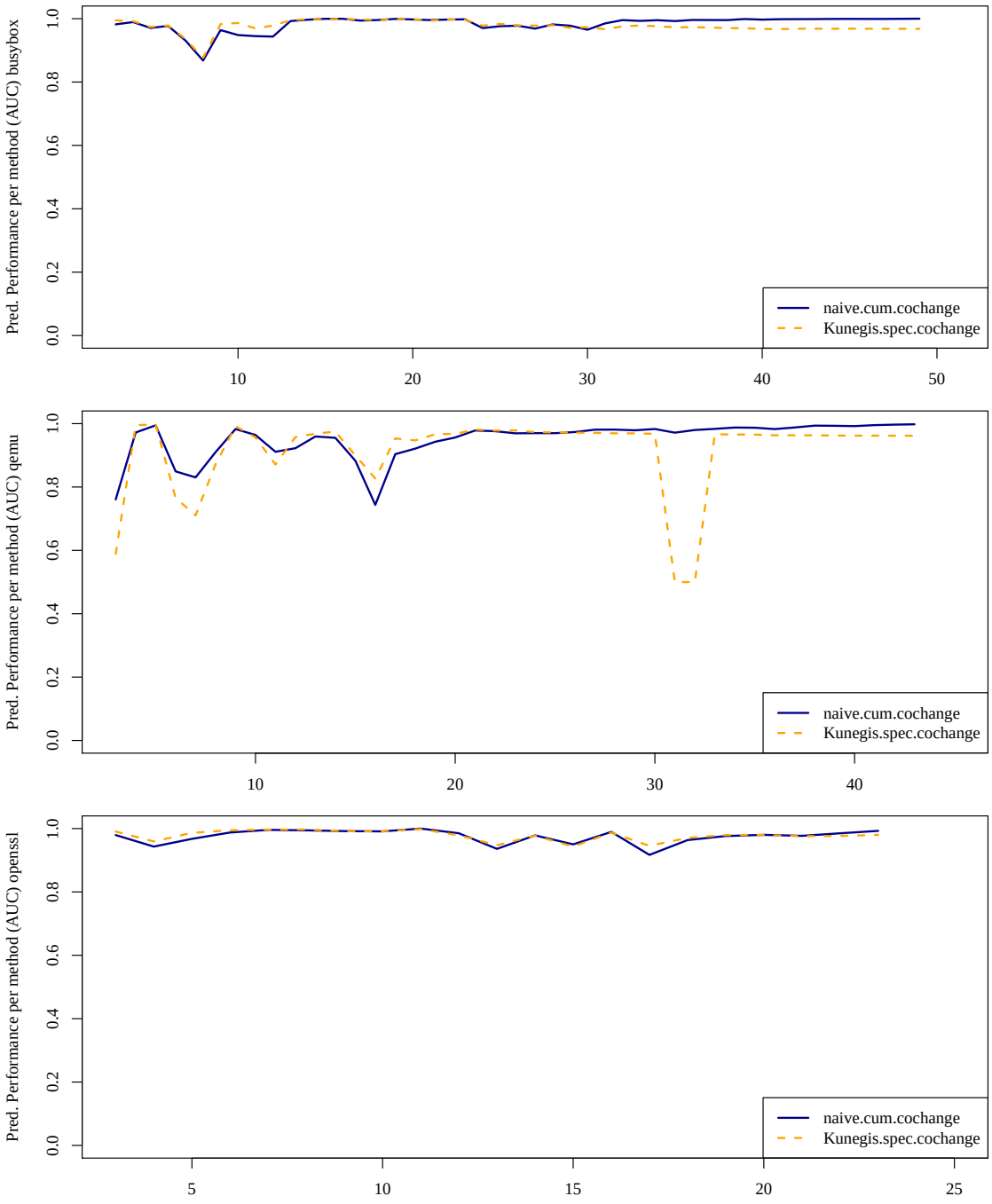

Figure 2.A.7: Predictive performance for analysis of cumulated `cochange` networks, for the prediction of the uncumulated adjacency matrices with information from the time t and $t-1$. Panel 1 refers to `BusyBox`, panel 2 to `QEMU` and panel 3 to `OpenSSL`.

Chapter 3

Spatial Detrending revisited: Modelling Local Trend Patterns in NO_2-Concentration in Belgium and Germany

In coauthorship with Prof. Dr. Harry Haupt and Svenia Behm, Faculty of Business Administration and Economics, Chair of Statistics, University of Passau

Chapter Abstract

Short-term predictions of air pollution require spatial modelling of trends, heterogeneities, and dependencies. Two-step methods allow real-time computations by separating spatial detrending and spatial extrapolation into two steps. Existing methods discuss trend models for specific environments and require specification search. Given more complex environments, specification search gets complicated by potential nonlinearities and heterogeneities. This research embeds a nonparametric trend modelling approach in real-time two-step methods. Form and complexity of trends are allowed to vary across heterogeneous environments. The proposed method avoids ad hoc specifications and potential generated predictor problems in previous contributions. Examining Belgian and German air quality and land use data, local trend patterns are investigated in a data driven way and are compared to results computed with existing methods and variations thereof. An important aspect of our empirical illustration is the heterogeneity and superior performance of local trend patterns for both research regions. The findings suggest that a nonparametric spatial trend modelling approach is a valuable tool for real-time predictions of pollution variables: it avoids specification search, provides useful exploratory insights and reduces computational costs.

3.1 Introduction

Industrial parks, roads and other sources of fossil fuel combustion processes are responsible for a large share of nitrogen oxides and particulate matters that pollute the air and create severe health risks (Wolf et al., 2017). Information on the location of pollution sources can enhance the identification of local pollution hotspots and trend patterns, even at points where no direct observations are available. Detailed spatial pollution maps have a considerable impact on health policy. An example is the German legislation on banning pollution-intensive cars from cities and its major impact on air pollution (Fensterer et al., 2014).

A well-established source of information for air quality assessment are land use classes. Land use data such as the CORINE land cover inventory encode the usage of a particular territory in land use classes (e.g., Feranec et al., 2016). Frequently, these classes are combined with complementary information on traffic density, demography, topography, and other geographic variables (e.g., Gilliland et al., 2005; Hooyberghs et al., 2006; Sahsuvaroglu et al., 2006; Janssen et al., 2008; Wang et al., 2013; Hennig et al., 2016). A key advantage of land use data is that information on single land use classes can be scaled down when granular data are available, for example on individual exposure to air pollution within a single urban residence (Hennig et al., 2016).

The crucial role of land use information in regression-based models has lead to the notion *Land Use Regression* (LUR). The difference between *using land use indicators in regression* and LUR is that the latter usually relies on the assumption of independence and stationarity of the regression errors (e.g., Gilliland et al., 2005; Ryan and LeMasters, 2007; Hoek et al., 2008). Neglecting such assumptions carries severe potential for ignoring bias and inefficiencies (Montero et al., 2015). Air pollution data are likely to exhibit spatial dependence, because the closer two monitoring sites are located, the more likely they share a common source of pollution or dominant wind direction. There are two main alternatives to combining a regression framework with the modelling of spatial dependencies among individual sites.

(a) In two-step or *residual kriging* methods, a first spatial detrending step allows

to filter nonstationarities driven by phenomena such as titration (e.g., Hooyberghs et al., 2006). This is followed by a second (ordinary) kriging step to include the dependence structure in the spatial prediction. Hooyberghs et al. (2006) and Janssen et al. (2008) suggest to use historical data to produce real-time spatial predictions within a two-step *residual interpolation optimised* (RIO) modelling framework. To account for nonstationarities in O_3-concentration across Belgium, Hooyberghs et al. (2006) compute a local spatial trend based on historical measurements using population density as auxiliary data. Janssen et al. (2008) use CORINE land use data instead of population density data in the detrending step and analyse the three pollutants NO_2, O_3, and PM_{10}. The RIO residual kriging procedure has two advantages: First, trend and semivariogram estimation can be done in two separated steps. Second, as long as the crucial assumption of stable spatial trend and semivariogram over time holds, it allows real-time predictions at basically zero computational cost.
(b) Alternatively, *universal kriging* is a one-step method, where the spatial dependence structure and the impacts of the predictors are estimated simultaneously. However, the difference between two-step methods and universal kriging is not always clear-cut (e.g., Mercer et al., 2011), and the latter can also be applied to filtered data. As Montero et al. (2015) point out, splitting up detrending and kriging in two steps is a recommended alternative to avoid ambiguities in universal kriging with regard to the interplay of trend specification and semivariogram estimation. While a correct trend specification is important in both methods to fulfill the requirements for kriging, it remains unclear how to specify the relationship between predictors and pollution with regard to optimising predictive performance.

Two-step methods provide a simple and useful tool for real-time predictions. Their key assumption seems to hold, as average pollution levels are quite stable over time and independent of short term influences, for example over different seasons (e.g., Sahsuvaroglu et al., 2006), or over the span of several years (e.g., Wang et al., 2013). Our work aims at providing further insights into two-step methods such as the RIO residual kriging method, and generalises the method of Janssen et al. (2008) theoretically and empirically. The quality of the trend filter in the first step is crucial for any inferences drawn from the second step. Hence we suggest nonparametric

generalisations to adapt the trend modelling step to general environments, exhibiting different degrees of complexity and heterogeneity in spatial patterns. In particular we suggest to simplify the inclusion of land use classes in the trend estimation step. In Janssen et al. (2008), every monitoring site is assigned a pollutant-specific land use indicator that describes average pollution based on the relative share of every land use class within the sites' vicinity. This indicator summarises the interplay of constant local characteristics contained in the predictors and is interpreted as a proxy for the long-term total pollution load a single location has to carry. The authors assume that mean and standard deviation of the pollutant can be described by polynomials in the indicator. They do not consider additional predictors controlling for further sources of heterogeneity in spatial trend patterns. To avoid the consequences of misspecifying the trends, we propose to use nonparametric trend regressions. Nonparametrics allow for a data-driven exploration of trend patterns while avoiding specification search based on ad hoc polynomials (and interactions if further predictors are used). We show that multivariate generalisations of the trend functions can be easily accomplished by allowing for different trends for background, industrial and traffic environments.

The simultaneous estimation of a trend function and a pollutant-specific land use indicator (weighting single land use classes) in prediction employed by Janssen et al. (2008) leads to a generated predictor problem. Hence we propose direct inclusion of the information on land use classes as predictors in our trend function. We thoroughly discuss estimation, prediction and comprehensive empirical evidence for Belgian and German air quality and land use data. Our empirical analysis reproduces existing results of Janssen et al. (2008) for Belgium and provides evidence for Belgium and Germany that the suggested modifications perform very well.

The remainder of this article is organised as follows: Section 2 discusses the database used for our empirical investigation. Section 3 explains the statistical theory, including an overview on Janssen et al. (2008) and indicator-based two-step spatial prediction methods. Section 4 provides detailed insights into our results and section 5 concludes.

3.2 Data

In the application to German air pollution, we investigate daily maxima of the recorded hourly NO_2-concentration over the time period 1st Jan 2007 to 31st Dec 2012. The data have been obtained from the European Environment Agency (EEA), who maintains AirBase, the European air quality database (EEA, European Environment Agency, 2016). The database consists of monitoring data from fixed monitoring sites, measured at regular intervals, as well as meta-information on the monitoring sites involved. One meta-information is the sites' type that can either be "Background", "Industrial", or "Traffic". For a complete description of the meta-data on monitoring site characteristics, we refer to 3.A.2. Further, we use the CORINE Land Cover 2006 (CLC2006) data layer in a 100×100 meter resolution (EEA, European Environment Agency, 2010b). For detailed information on CLC data including changes between the four different data layers CLC1990, CLC2000, CLC2006, CLC2012, see Feranec et al. (2016).

In order to make our empirical findings comparable to those of Janssen et al. (2008), we also analyse Belgian hourly NO_2-concentration from AirBase over the time period 1st Jan 2001 to 31st Dec 2006, and the CLC2000 layer, i.e. land use classification in the year 2000 version (EEA, European Environment Agency, 2010a). Table 3.1 shows that German data contain a considerably higher number of monitoring sites and exhibit a quite different distribution over measuring sites' types in comparison to Belgium. While both countries have an equivalent share of background sites, the relative shares of industrial and traffic sites are inverted.

Table 3.1: Numbers of monitoring sites in Belgium (Germany) that were active within the period 1st Jan 2001 to 31st Dec 2006 (1st Jan 2007 to 31st Dec 2012).

	Background	Industrial	Traffic	Total
Belgium	37 (52.85%)	23 (32.86%)	10 (14.29%)	70
Germany	276 (51.49%)	38 (7.09%)	222 (41.42%)	536

In our analysis we omit daily maximum NO_2 values above 500 $\mu g/m^3$ as well as negative values. Based on the remaining daily maximum values the mean and standard deviation of each monitoring site is calculated, separately for weekdays and

weekends. For supplementary information about the data quality of the German and Belgian air pollution data and the data preprocessing we refer to 3.A.2. Fig. 3.1 displays the respective boxplots for Belgium and Germany. While the four statistics

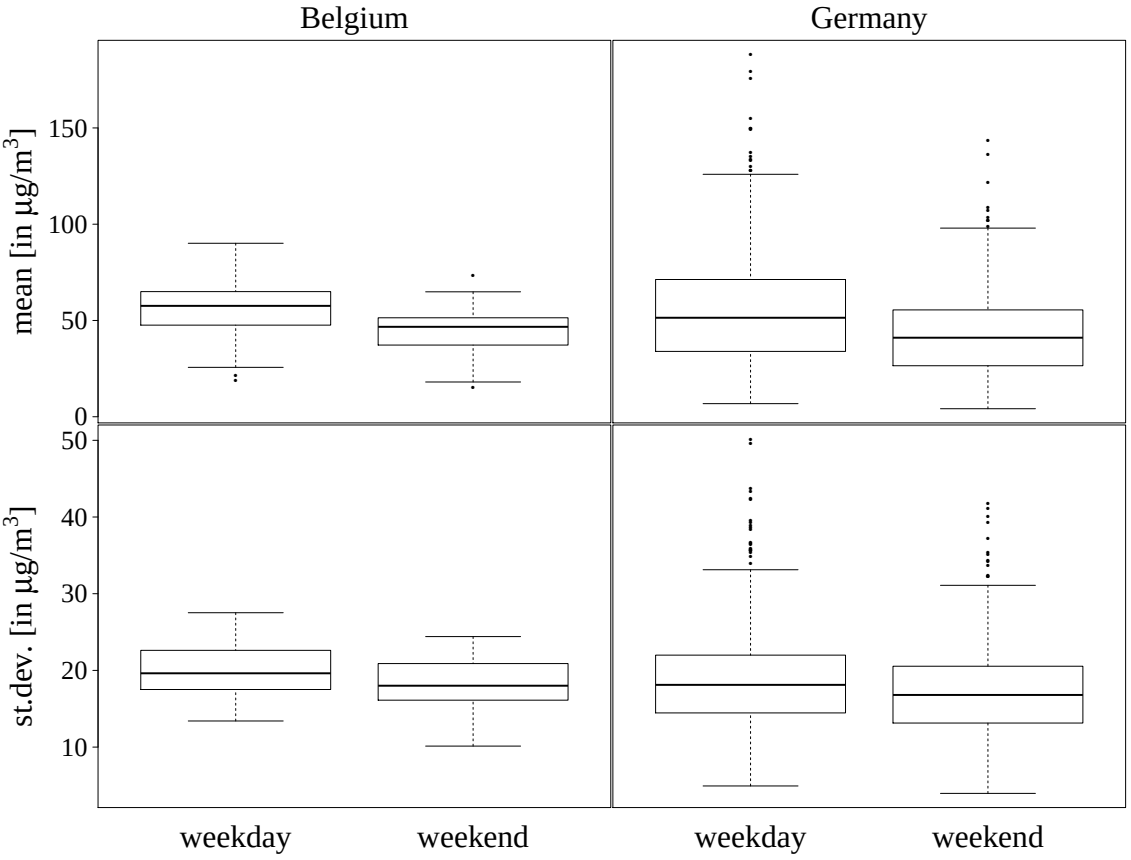

Figure 3.1: Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian monitoring site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

(mean weekday, mean weekend, st.dev. weekday, st.dev. weekend) for Belgium and Germany differ only slightly with respect to their medians, the interquartile ranges and the ranges between the whiskers are remarkably higher for the German data compared to Belgian data. For both research regions we observe differences between the mean of daily maximum NO_2 concentrations on weekdays and weekends. For the standard deviation of daily maximum NO_2 concentrations only a small difference between weekdays and weekends occurs. In Figs. 3.A.1-3.A.3 we explore the distribution of the means and standard deviations differentiating by the sites' type. We find that observed differences between Belgium and German data can be traced

back to measurements at traffic sites.

Considering the usage of the CLC data in air pollution studies, it is common practice to reclassify the 44 land use classes in the CLC inventory (e.g. Beelen et al., 2009, 2013; Wolf et al., 2017). Following the suggestion of Janssen et al. (2008), we group the 44 classes into eleven more general land use classes. The European Monitoring and Evaluation Programme (EMEP) provides emission data concerning national total, sector and gridded emissions for Europe (see EMEP and CEIP, 2014, for detailed information). Those data are classified with regard to their relationship to air pollution, and the classification results in so-called sectors, referred to as SNAP (Selected Nomenclature for reporting of Air Pollutants). Table 3.2 summarises the resulting classifications and descriptions.

Table 3.2: Relationship between grouped CLC classes and the equivalent groups in the SNAP sector classification (according to Janssen et al., 2008).

grouped class	**description**	**CLC classes**	**SNAP sectors**
class 1	Continuous urban fabric	1	S2
class 2	Discontinuous urban fabric, green and sport	2,10,11	S2
class 3	Industrial or commercial units	3	S3+S4
class 4	Road and rail networks and associated land	4	S7
class 5	Port areas	5	S8
class 6	Airports	6	S8
class 7	Mine, dump and construction sites	7-9	S1+S4+S5+S9
class 8	Arable land	12-14	S10
class 9	Agricultural areas	15-22	S10
class 10	Forest and semi natural areas	23-34	S11
class 11	Wetlands and water bodies	35-44	S11

The empirical analysis is conducted with the statistical software `R` (`R` Core Team, 2013) using the packages broom (Robinson, 2017), GISTools (Brunsdon and Chen, 2014), gstat (Pebesma, 2004; Gräler et al., 2016), np (Hayfield and Racine, 2008), optimx (Nash and Varadhan, 2011; Nash, 2014), raster (Hijmans, 2016), rgdal (Bivand et al., 2017), spatstat (Baddeley et al., 2015), and timeDate (`Rmetrics` Core Team et al., 2015).

3.3 Statistical modelling

Assume air pollution at time $t \in D_t$ to be a latent geostatistical random process

$$Y_t(\cdot) = \{Y_t(\boldsymbol{s}) : \boldsymbol{s} \in D_s \subset \mathbb{R}^2\},$$

where D_s refers to the study area. Within the study region D_s define the locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, $n \in \mathbb{N}$. Let $Z_t(\boldsymbol{s})$, where

$$Z_t(\cdot) = \{Z(\boldsymbol{s}, t) : \boldsymbol{s} \in D_s\},$$

denote the data process at time $t \in D_t$. In our computations below let $z_{i,t}$ denote a realisation of $Z_t(\boldsymbol{s}_i)$ at location $\boldsymbol{s}_i$ at time $t \in D_t$. The vector $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,T}), i \in 1, \ldots, N$, defines the time series at monitoring site i, the vector $\boldsymbol{z}_t = (z_{1,t}, \ldots, z_{n,t}), t \in 1, \ldots, T$, defines measurements for all cross-sectional units or monitoring sites recorded at time t.

Following Cressie (1993) and Diggle and Ribeiro Jr (2007), the relationship between the unobserved geostatistical process and the data process is given by

$$Z_t(\boldsymbol{s}) = Y_t(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s}) \tag{3.1}$$

with $\epsilon_t(\boldsymbol{s}) \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. If the unobserved geostatistical process $Y_t(\cdot)$ at time $t \in D_t$ is assumed to be a stationary and isotropic Gaussian process, it holds $\forall \boldsymbol{s}, \boldsymbol{s'} \in D_s$, $\boldsymbol{s} \neq \boldsymbol{s'}$,

$$E[Y_t(\boldsymbol{s})] = \mu, \tag{3.2a}$$

$$Var[Y_t(\boldsymbol{s})] = \sigma^2, \tag{3.2b}$$

$$C(h) = Cov[Y_t(\boldsymbol{s}), Y_t(\boldsymbol{s}')] = \sigma^2 \rho(h), \tag{3.2c}$$

where the autocorrelation function $\rho(h) = Corr[Y_t(\boldsymbol{s}), Y_t(\boldsymbol{s}')]$ depends on the distance $h = ||\boldsymbol{s} - \boldsymbol{s}'||$, $E[\cdot]$ denotes the expected value, $Var[\cdot]$ the variance, and $C(\cdot)$ the autocovariance function. Under the assumptions stated above, analogous stationarity conditions hold for the data process $Z_t(\cdot)$, and the ordinary kriging predictor

$\hat{Y}_t(\boldsymbol{s}_0)$ can be calculated for any $\boldsymbol{s}_0 \in D_s, t \in D_t$.

3.3.1 Spatial Trend Modelling: Parametric Polynomials

The RIO technique proposed by Hooyberghs et al. (2006) and Janssen et al. (2008) starts with a detrending step in order to filter the data process $Z_t(\cdot)$ such that stationarity conditions analogous to (3.2a)-(3.2c) hold. The grouped land use classes (see Table 3.2) enter the equation for the pollutant specific β-index according to

$$\beta(\boldsymbol{s}, r) = log\left[1 + \sum_{k=1}^{11} a_k \cdot sh_k(\boldsymbol{s}, r)\right], \tag{3.3}$$

where $sh_k(\boldsymbol{s}, r)$ describes the share of the k-th class within a circular buffer zone with radius r around location $\boldsymbol{s}$. For the sake of simplicity we omit r and $\boldsymbol{s}$ and write β_i for $\beta(\boldsymbol{s}_i, r)$, β for $\beta(\boldsymbol{s}, r)$ and sh_k for $sh_k(\boldsymbol{s}, r)$. The class weights a_k, $k = 1, \ldots, 11$, define the relative impact of the respective class on the concentration of the air pollutant under investigation. Eq. (3.3) shows how the relative contribution of every land use class is summed up to an overall indicator. This means that a certain share of roads can be equivalent to a certain share of industrialised area, or a larger share of residential area (as the latter are usually relatively small sources of air pollution). Further details on the class weights are given in Table 3.A.1 in the Appendix.
Janssen et al. (2008) assume that spatial trends of mean and standard deviation are functions of the pollutant specific β-index. For the sake of a more general exposition covering the extensions in Section 3.2, we consider trend functions including potential further predictors X,

$$\mu \approx m_\mu(\beta, X), \tag{3.4a}$$

$$\sigma \approx m_\sigma(\beta, X). \tag{3.4b}$$

In their application to Belgian data, Janssen et al. (2008) assume that mean and standard deviation in Eqs. (3.4a) and (3.4b) can be described by a second and first order polynomial of β, respectively, and do not consider additional predictors X. The functions m_μ and m_σ are estimated in regressions using estimates $\bar{z}$ and s of μ

and σ, respectively, based on the time series observed for each measuring site where a distinction is made between weekdays and weekends. For the sake of simplicity we omit further notation.

For both statistics, β is calculated via Eq. (3.3) and therefore depends on $\boldsymbol{s}$ and $a_1, \ldots, a_{11}$. Under assumption (3.4a) the coefficients $a_1, \ldots, a_{11}$ in Eq. (3.3) are optimised through the following numerical optimisation procedure, after defining suitable termination criteria

1. Specify a starting set $a_1^{(1)}, \ldots, a_{11}^{(1)}$ of $a_1, \ldots, a_{11}$ (see Janssen et al., 2008).
2. Regress $\bar{z}_i$ on $m_\mu(\beta_i^{(1)}, X_i)$ where $\beta^{(1)}$ is computed using the set $a_1^{(1)}, \ldots, a_{11}^{(1)}$, and obtain the predictor $\widehat{m_\mu}^{(1)}(\beta_i^{(1)}, X_i)$.
3. Calculate the value of the RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{m_\mu}^{(1)}(\beta_i^{(1)}, X_i) - \bar{z}_i)^2}$.
4. If none of the termination criteria is fulfilled, restart the procedure with a different set $a_1^{(2)}, \ldots, a_{11}^{(2)}$, otherwise the optimal set is found.

Denoting the optimised class weights by $\tilde{a}_1, \ldots, \tilde{a}_{11}$ and the corresponding β-index by $\tilde{\beta}_1, \ldots, \tilde{\beta}_n$, the trend functions for mean and standard deviation can be computed, for every i, as $\hat{\mu}_i = \widehat{m_\mu}(\tilde{\beta}_i, X_i)$ and $\hat{\sigma}_i = \widehat{m_\sigma}(\tilde{\beta}_i, X_i)$, respectively.

According to Janssen et al. (2008), using the fitted values $\hat{\mu}_i$ and $\hat{\sigma}_i$, and given predefined reference levels μ^{ref} and σ^{ref}, detrending of the measurement values $z_{i,t}$ can be achieved according to

$$z_{i,t}^{*} = z_{i,t} + (\mu^{ref} - \hat{\mu}_i), , \tag{3.5a}$$

$$z_{i,t}^{**} = (z_{i,t}^{*} - \bar{z}_i^{*})\frac{\sigma^{ref}}{\hat{\sigma}_i} + \bar{z}_i^{*}. \tag{3.5b}$$

After filtering the monitored data $z_{i,t}$ according to Eqs. (3.5a) and (3.5b), we obtain the transformed data $z_{i,t}^{**}$, which we interpret as realisations of $Z_t^{**}(\boldsymbol{s}_i)$, the filtered data process at time $t \in D_t$. Hence, for each $\boldsymbol{s} \in D_s$,

$$E[Z_t^{**}(\boldsymbol{s})] = \mu(\boldsymbol{s}) + (\mu^{ref} - \hat{\mu}(\boldsymbol{s})) \approx \mu^{ref}, \tag{3.6}$$

relying on assumption (3.4a) in the last transformation, and

$$Var[Z_t^{**}(\boldsymbol{s})] = \left(\sigma^{ref}\right)^2 Var\left[\frac{Z_t^*(\boldsymbol{s}) - \bar{Z}^*(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}\right] \approx \left(\sigma^{ref}\right)^2, \tag{3.7}$$

since the middle term describes the standardisation of $Z_t^*(\boldsymbol{s})$. Eqs. (3.6) and (3.7) show that the filtered data process approximately satisfies the (weak) stationarity properties (3.2a)-(3.2c) and can be used in the kriging procedure.
Based on all historical detrended measurements $z_{i,t}^{**}$, the semivariogram required for ordinary kriging is estimated. For any $\boldsymbol{s}_0 \in D_s$ at time $t \in D_t$ an interpolated value $\hat{Y}_t^{**}(\boldsymbol{s}_0)$ can be calculated and retrended with regard to the local mean and local standard deviation of the originally monitored process. The retrending formulas can be written as

$$\hat{Y}_t^*(\boldsymbol{s}_0) = (\hat{Y}_t^{**}(\boldsymbol{s}_0) - \bar{\hat{Y}}^{**}(\boldsymbol{s}_0))\frac{\hat{\sigma}(\boldsymbol{s}_0)}{\sigma^{ref}} + \bar{\hat{Y}}^{**}(\boldsymbol{s}_0), \tag{3.8a}$$

$$\hat{Y}_t(\boldsymbol{s}_0) = \hat{Y}_t^*(\boldsymbol{s}_0) - (\mu^{ref} - \hat{\mu}(\boldsymbol{s}_0)). \tag{3.8b}$$

The RIO technique rests on the crucial assumption that both spatial trends and the semivariogram are stable over time, enabling real-time predictions at basically zero computational cost. Real-time predictions are produced in the following way: detrend a new set of observations (at monitoring sites) using the fitted trend functions, interpolate the detrended values using the fitted semivariogram and retrend the interpolated values using the fitted trend functions.

3.3.2 Spatial Trend Modelling: a General Nonparametric Approach

There are several options to include further predictors in Eqs. (3.4a) and (3.4b). In general, the functions m_μ and m_σ can be approximated by higher-order parametric expansions using polynomials of β interacting with (the levels of) X. Such a strategy, however, requires assumptions on the degree of the approximation and a high number of parameters. In order to avoid ad hoc assumptions, potential underspecification, or potentially extensive specification search, a straightforward alternative is to estimate m_μ and m_σ using a nonparametric trend model. Such a model should

deliver a more accurate representation of the trend patterns than a specification based on a parametric expansion if the latter is underspecified and the data are sufficiently informative for nonparametric regression (e.g., Haupt et al., 2010). More important and evident from our empirical illustration, nonparametric methods provide explorative insights about the trend patterns driven by β and potential further predictors such as the type of monitoring sites X.

Hence, as nonparametric methods can help to identify the best parametric approximation and to avoid problems of misspecifying the trend functions, we employ a local linear kernel smoothing estimator of $E(\bar{Z}|\beta, X) = m(\beta, X)$ in the trend regression model

$$\bar{Z} = m(\beta, X) + U \qquad \text{with } E(U|\beta, X) = 0, \tag{3.9}$$

based on Eq. (3.4a). A generalised least squares estimator is denoted as $\widehat{m}_{LL}$, where $(\widehat{m}_{LL}, \widehat{\gamma})$ minimises

$$\sum_{i=1}^{n} [\bar{Z}_i - m - \gamma(\beta_i - \beta)]^2 K(\mathbf{W}, \mathbf{W}_i, \mathbf{h}),$$

where $\mathbf{W} = (\beta, X)$ denotes the vector of regressors, $K = k_\beta \cdot k_X$ is a product kernel, and $\mathbf{h} = (h_\beta, h_X)'$ is a vector of bandwidths which we estimate using least squares cross validation (see Li and Racine, 2004). The use of *mixed* continuous (i.e. pollutant specific β-index) and categorical (i.e. type of monitoring site X) predictors in nonparametric regressions has been discussed extensively in the various works of Li and Racine (e.g., Li and Racine, 2007).

The β-index in Eq. (3.9) is unknown and has to be computed according to the procedure described in section 3.1. Hence the estimated β-index $\tilde{\beta}$ is a generated predictor. The potential consequences for estimation and inference in parametric models have been discussed in an abundant literature following the seminal paper of Pagan (1984). In a nonparametric context Sperlich (2009) and Mammen et al. (2012) provide authoritative treatments (see Haupt et al., 2018, for a discussion in the mixed predictor context). Depending on the problem at hand, researchers may prefer to use an aggregated index, but should be aware that generated regressor problems may invalidate the interpretation of the β-index. In the current context the problems can

be avoided from the outset if the β-index is not considered. We propose to directly include the information on land use classes and define the categorical predictors

$$X_1 = \underset{k\in\{1,\ldots,11\}}{\operatorname{argmax}} sh_k, \tag{3.10}$$

$$X_2 = \underset{k\in\{1,\ldots,11\}\setminus X_1}{\operatorname{argmax}} sh_k, \tag{3.11}$$

determining which classes have the largest and second largest share (within the circular buffer zone around a certain location), respectively. Note that including the third largest class has no remarkable effect. In our application for 534 of 536 German sites and for 69 of 70 Belgian sites, the sum of the shares of the first and second largest class is larger than 50%. The continuous predictor

$$S = sh_{X_1} + sh_{X_2}. \tag{3.12}$$

is defined as the sum of the shares of the first and second largest class. Then, instead of the predictors used in Eq. (3.9), we consider the categorical predictors X_1, X_2, and the sites' type, and the continuous predictor S.

3.4 Results

For the sake of exposition, we introduce the following abbreviations: "QL" ("LL") refers to a quadratic (linear) trend for the mean and a linear trend for the standard deviation; "TypeQL" ("TypeLL") allows local trend differing with respect to the type of a monitoring site ("Background", "Industrial", or "Traffic"); "NP" refers to the nonparametric approach with $\beta-$index together with the sites' type; "NPnoBeta" refers to the nonparamatric approach without $\beta-$index but with the predictors defined in Eqs. (3.10)-(3.12) together with the sites' type.

The estimated QL trend functions for Belgian data are displayed in Fig. 3.2 and replicate results of Janssen et al. (2008, right plot of Fig. 5 and middle plot of Fig. 8), while Fig. 3.3 shows the corresponding QL estimates using German data. A global second order polynomial fits the Belgian data quite well, while we ob-

serve considerably more heterogeneity in the German data. The curvature is less pronounced in the plots of Fig. 3.3 and bear no visible difference to the LL trend functions displayed for Germany in Fig. 3.A.8 in 3.A.1.

Comparing the trend functions for weekdays and weekends for Belgian as well as for German data, we observe a shift along the y-axis (for both specifications LL and QL). This is in accordance with the boxplots displayed in Fig. 3.1 and indicates that, on average, the concentration level of NO_2 drops from weekdays to weekends.

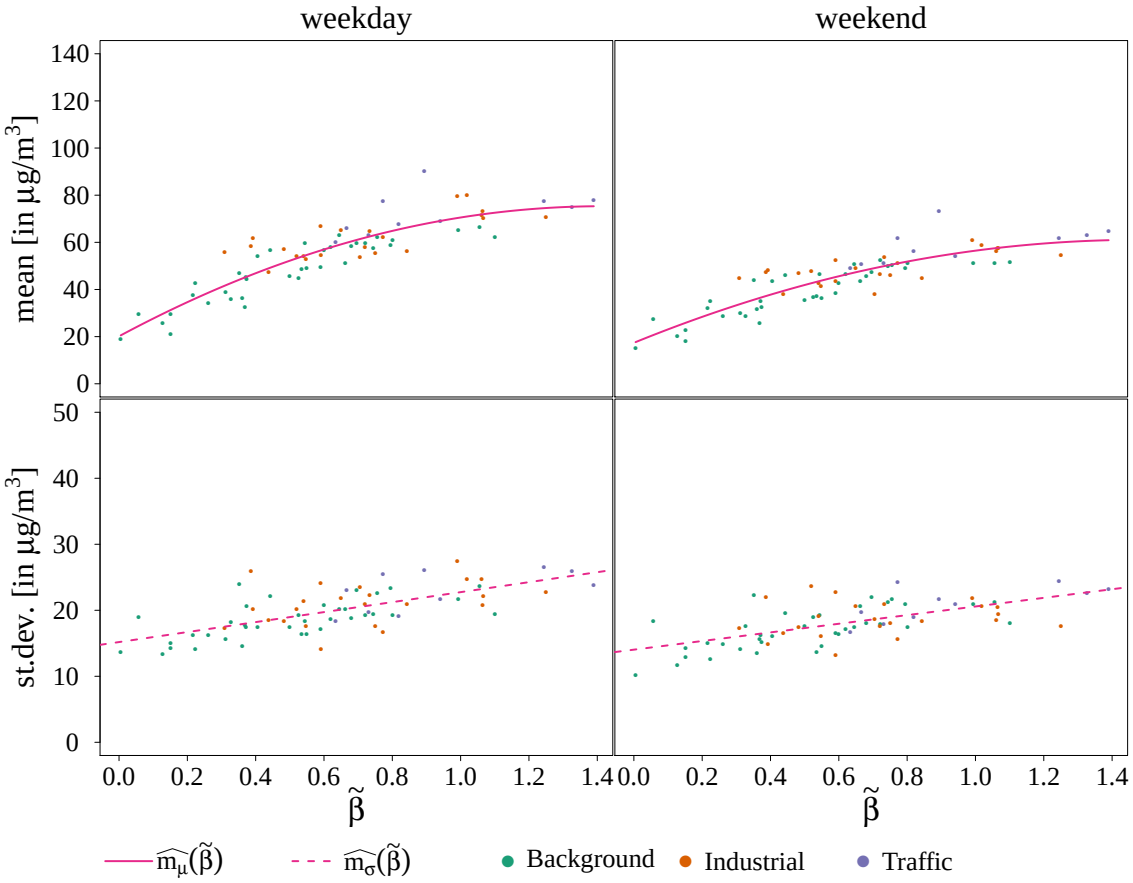

Figure 3.2: Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation (specification QL).

The replication of the results of Janssen et al. (2008) in a narrow sense for Belgian data and in a wider sense for German data suggests that the assumption of global trend forms is too restrictive. Determining a global trend form requires an ad hoc specification of polynomial degree and specification search. Previous contributions such as Janssen et al. (2008) do not explicitly discuss this issue. The optimisation

of the class weights a_k affects the values of $\tilde{\beta}$, the position of the points along the x-axis and thus the fitted trend function (e.g., compare the range of $\tilde{\beta}$ in Fig. 3.3 and Fig. 3.A.8). To avoid ad hoc specification search and to widen the scope of applicability to heterogeneous environments, we discuss a more general approach to spatial trend fitting and illustrate it with German data. Note that further results for Belgium, completing our empirical analysis, are provided in 3.A.1.

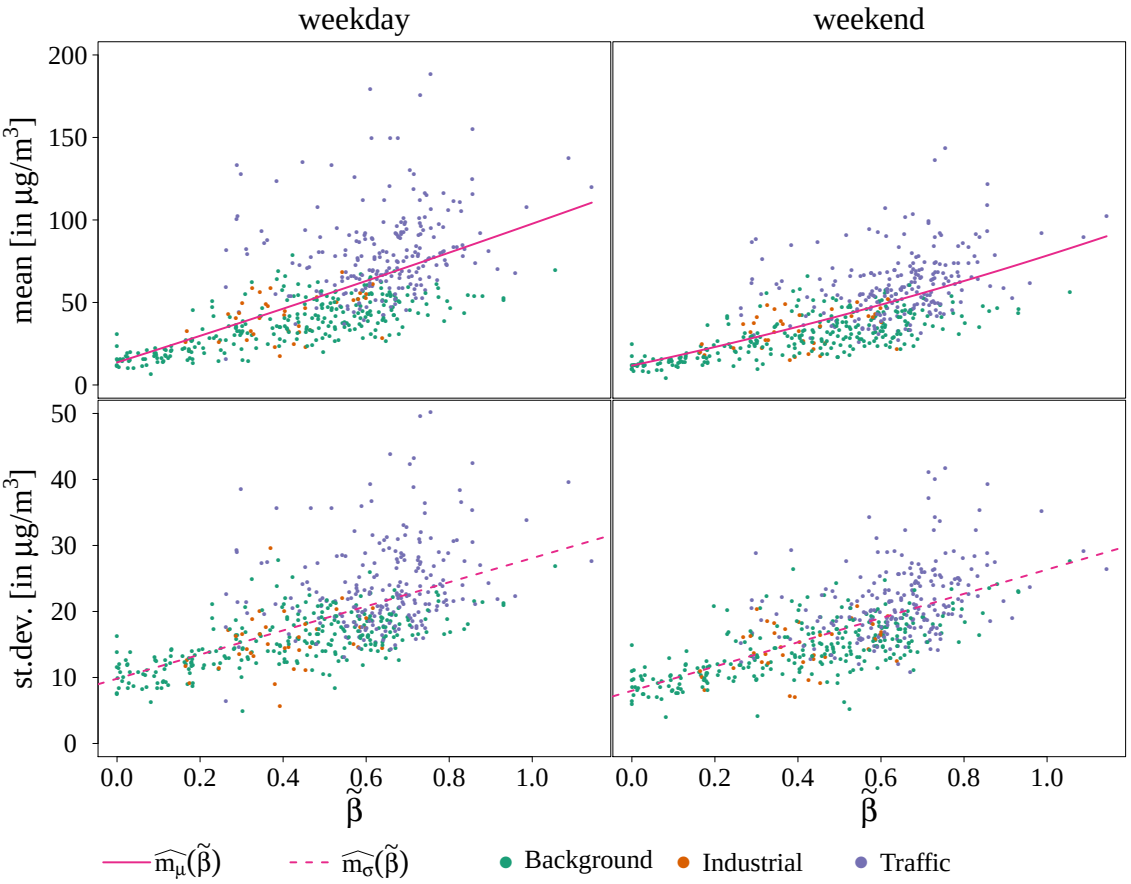

Figure 3.3: German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation (specification QL).

An encompassing approach to trend analysis is the nonparametric regression, following the mixed kernel estimation approach for continuous and categorical predictors of Li and Racine (2004, 2007), compare Eq. (3.9). Fig. 3.4 shows estimated NP trend functions for German data based on local linear kernel regressions, where bandwidths are estimated by least squares cross-validation using the default kernel functions proposed by Hayfield and Racine (2008). Trends are calculated by simul-

taneously smoothing over $\tilde{\beta}$ and the three categories of the sites' type contained in X. We observe substantial differences in local levels and slopes between traffic

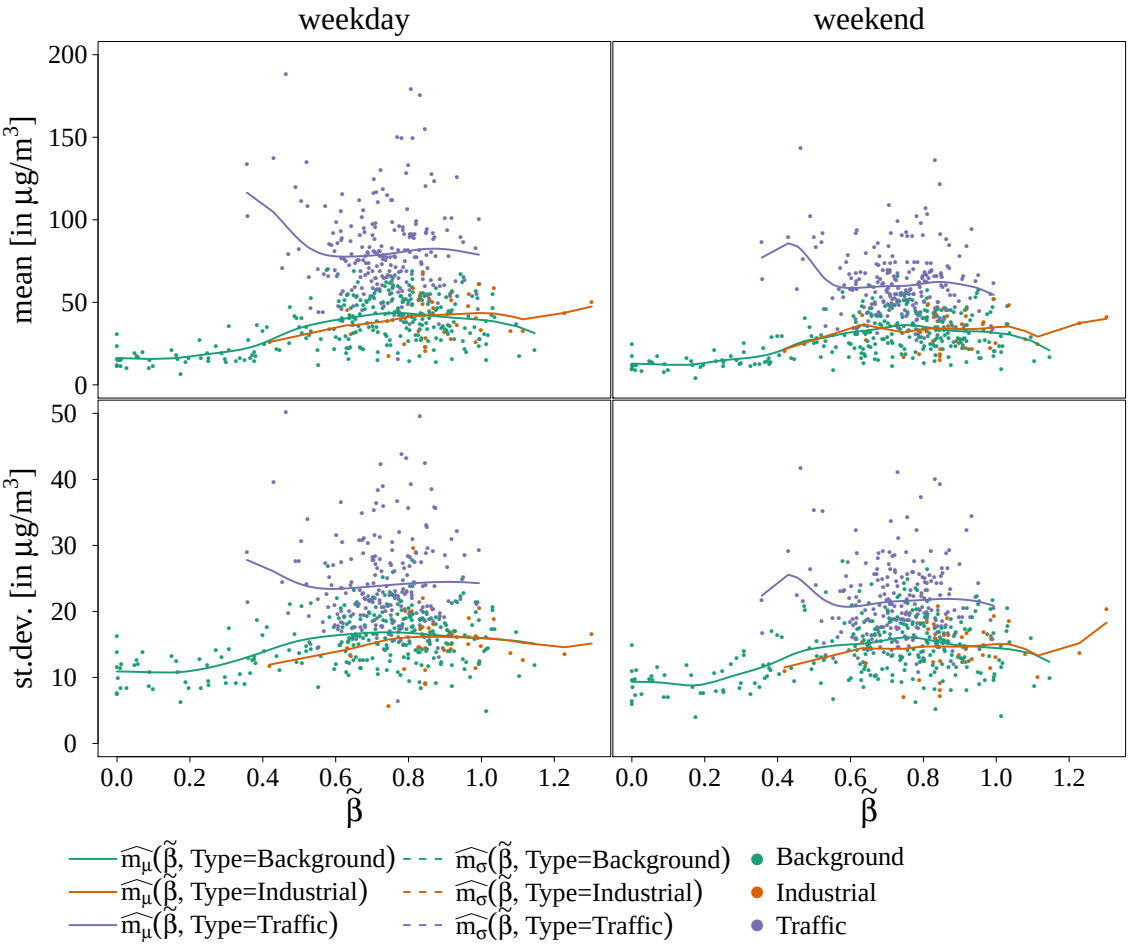

Figure 3.4: German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to the nonparametric approach (specification NP).

sites and all other sites indicating that the NO_2 concentration at traffic sites is on average larger than at background or industrial sites. Apart from minor boundary effects visible in the plots for weekend data, the estimates suggest that a piecewise quadratic trend may be sufficiently flexible. The finding of heterogeneity in local trend patterns in Germany based on our visual analysis is confirmed by the quantitative results from the nonparametric approach including the sites' type. The corresponding results on predictive performance are discussed in detail below.

Based on the exploratory insights obtained from the nonparametric regressions, we add dummy variables and interactions as indicators for the monitoring sites'

type to the specification QL. The resulting TypeQL trend estimates are shown in Fig. 3.5. Visual inspection of the results and comparison to Fig. 3.3 suggest that the

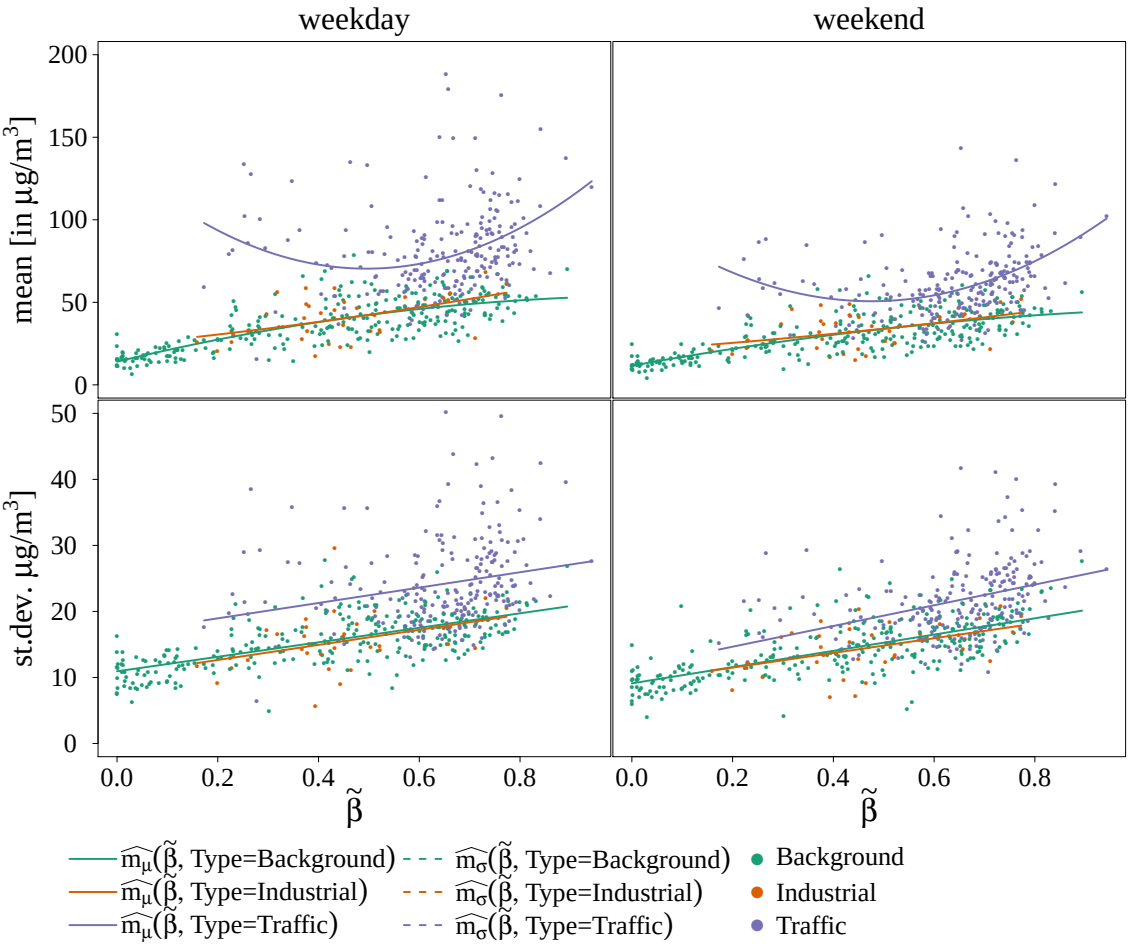

Figure 3.5: German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeQL).

specification TypeQL allowing local quadratic trend patterns provides a superior fit to the German data. Again, this finding is supported by an analysis of predictive performance. Equivalent plots for specifications LL and TypeLL for Germany are provided in Figs. 3.A.8 and 3.A.9 in 3.A.1.

The trend functions corresponding to the specifications TypeQL and TypeLL reveal substantial differences in local levels and slopes between traffic sites and all other sites in Germany. For Belgian data such clear differences cannot be observed (see Figs. 3.A.6 and 3.A.7 in 3.A.1).

For specification NPnoBeta trends are calculated by simultaneously smoothing over S and the categories X_1, X_2, and the sites' type. This specification entails considerably lower computational costs compared to those of NP, as the optimisation of group weights is not required. For German (Belgian) data computation time equals 3.45 hours (7 minutes) to derive the trend functions using NP, compared to 16 seconds (4.3 seconds) for NPnoBeta. For NPnoBeta it is not possible to display the estimated trend functions in two-dimensional space, as they depend on one continuous and three unordered categorical predictor variables. In order to evaluate the predictive performance of NPnoBeta compared to the approaches including the β-index, we carry out a leave-one-out cross-validation (LOOCV). In each loop of LOOCV one monitoring site is omitted and the entire RIO technique – consisting of the four steps of optimising group weights, detrending, kriging and retrending (as described in Section 3 above) – is applied to the remaining sites. For NPnoBeta the optimisation of group weights is no longer necessary and therefore each loop of LOOCV consists of the steps detrending, kriging and retrending. Table 3.3 summarises the results of LOOCV. As suggested by our visual inspection of the nonparametric trend estimates, allowing the trend functions to differ with the sites' type enhances the predictive performance. Adding an indicator for the sites' type to specifications QL (LL) leads to a performance gain of 13.7% (12.5%) with regard to RMSE for Germany. For Belgium, it lowers the RMSE by 2.0% when the indicator is added to LL, and increases the RMSE by 14.0% when the indicator is added to QL. The latter deterioration of predictive performance in Belgium is due to a single outlier produced in the optimisation process. Avoiding the generated predictor problem by including the information on land use classes directly in NPnoBeta improves (reduces) the predictive performance by 3.4% (1.6%) for German (Belgian) data compared to NP. Table 3.A.2 in 3.A.1 provides further and more detailed results on our LOOCV analysis, revealing that the inclusion of the third largest LUC class has no remarkable effect on the predictive performance with regard to RMSE. Overall we observe that NPnoBeta has a superior (equal) LOOCV performance for Germany (Belgium) while it does not require specification search, avoids generated predictor problems and causes almost zero computational costs.

Table 3.3: Results of LOOCV for different specifications and their predictive performance.

RMSE	QL	LL	TypeQL	TypeLL	NP	NPnoBeta
Germany	20.84	20.82	17.99	18.21	19.07	18.43
Belgium	13.76	13.79	15.69	13.51	13.66	13.88

3.5 Discussion and Conclusions

Approaches for spatial interpolation of air pollutant data require assumptions on stationarity or on trend patterns of the underlying geostatistical random processes. Step-wise procedures based on filtering known or estimated spatial trends bear the advantage of real-time applicability due to their computational and interpretational simplicity. The RIO framework of Hooyberghs et al. (2006) and Janssen et al. (2008) enhances spatial interpolation and predictive performance by exploiting pollution relevant information from local land use patterns. The general applicability of the method hinges on assumptions about ad hoc global trend patterns defined by land use related pollution indicators. Existing methods discuss trend models for specific environments and require specification search. In practice, however, research environments of different size and level of aggregation may exhibit complex nonlinear local trend patterns, driven by spatial heterogeneities and dependencies. Specification search then becomes a troublesome endeavour.

Based on the spatial detrending employed by Janssen et al. (2008), we propose the use of a simple flexible framework for data driven trend modelling and subsequent filtering of the data. A crucial assumption is the selection of further predictors driving the spatial complexity of trend patterns. The various types of monitoring sites are an obvious initial choice for such a predictor. This approach has the advantage of preserving the intuition of larger values of the land use indicator β representing higher local – that is type-specific – levels of pollution, while allowing for type-specific trend levels and slopes.

We propose a nonparametric spatial trend modelling approach using all available predictors. The approach is computationally feasible and does not require ad hoc assumptions on the functional form. It can be used in an exploratory way to identify

potential parametric approximations of trend generating mechanisms. In addition, we propose to avoid potential generated predictor problems. This can be done by directly including the information on land use classes, instead of computing a pollution-specific indicator. The performance of the proposed method, existing methods, and variants thereof can be studied by using leave-one-out cross-validation analysis of the predictive performance.

We find that a simple generalisation of the existing methods by using multiple non-parametric regression methods leads to considerable gains in predictive performance while computational costs remain low. Furthermore, the proposed method bears a large potential for exploratory analysis of trending mechanisms while avoiding lengthy trend specification search.

In an empirical study, we first successfully replicate existing results of Janssen et al. (2008) for Belgium using similar but not the same data, and then apply the proposed method to German data. We investigate the assumption of global trend patterns and find strong (weak) evidence against such an assumption for German (Belgian) data. The nonparametric approach can be used to identify local parametric approximations of trend patterns. The overall performance of the proposed method suggests that the nonparametric method is a very good choice for research environments with considerably different complexity. Obvious advantages are that it does not require specification search, avoids generated predictor problems and has almost zero computational costs.

Potential extensions can be considered in several directions. First, it should be kept in mind that the β-values change simultaneously with the functional form, and hence a monotonicity restriction is necessary to preserve the intuition of β as an index representing mean pollution. A non-monotonic functional form resulting from polynomial or nonparametric trend fits stresses plausibility of this theoretical rationale. The question of imposing monotonicity constraints or not depends on the problem at hand; i.e. whether predictive performance or interpretability is the main objective. Second, statistical tools could be used to provide live monitoring of the crucial assumption of stable trend functions for mean and standard deviation over time. Third, the robustness of the results could be assessed with regard to the choice

and aggregation of land use categories as well as the choice of variables determining the trend forms. Fourth, further diagnostics could refer to the uncertainty arising from the stepwise nature of the analysis. There is no clear indication in the original application on how to calculate the uncertainty arising from errors due to trend elimination and kriging, as well as their potential dependence structure.

A flexible two-step procedure reduces the computational demand for spatial now- and forecasts and allows researchers to explore and test suitable trend specifications. The approach is transparent in its single steps and sufficiently general for a wide range of applications.

3.6 References

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press, London, UK.

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., Briggs, D.J., 2009. Mapping of background air pollution at a fine spatial scale across the European Union. Science of the Total Environment 407, 1852–1867. doi:10.1016/j.scitotenv.2008.11.048.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M., Künzli, N., Schikowski, T., Marcon, A., Eriksen, K.T., Raaschou-Nielsen, O., Stephanou, E., Patelarou, E., Lanki, T., Yli-Tuomi, T., Declercq, C., Falq, G., Stempfelet, M., Birk, M., Cyrys, J., von Klot, S., Nádor, G., Varró, M.J., Dėdelė, A., Gražulevičienė, R., Mölter, A., Lindley, S., Madsen, C., Cesaroni, G., Ranzi, A., Badaloni, C., Hoffmann, B., Nonnemacher, M., Krämer, U., Kuhlbusch, T., Cirach, M., de Nazelle, A., Nieuwenhuijsen, M., Bellander, T., Korek, M., Olsson, D., Strömgren, M., Dons, E., Jerrett, M., Fischer, P., Wang, M., Brunekreef, B., de Hoogh, K., 2013. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe – the ESCAPE project. Atmospheric Environment 72, 10–23. doi:10.1016/j.atmosenv.2013.02.037.

Bivand, R., Keitt, T., Rowlingson, B., 2017. rgdal: Bindings for the Geospatial Data Abstraction Library. URL: https://CRAN.R-project.org/package=rgdal. R package version 1.2-7.

Brunsdon, C., Chen, H., 2014. GISTools: some further GIS capabilities for R. URL: https://CRAN.R-project.org/package=GISTools. R package version 0.7-4.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley Series in Probability and Statistics. revised ed., Wiley, New York, USA.

Diggle, P.J., Ribeiro Jr, P.J., 2007. Model-based Geostatistics. Springer, New York, USA. doi:10.1007/978-0-387-48536-2.

EEA, European Environment Agency, 2010a. CORINE land cover 2000 raster data, version 13 (05/2010). URL: https://www.eea.europa.eu/ds_resolveuid/b00116e51c79865cf89a84162b8fd21e. Acessed on 29th May 2017.

EEA, European Environment Agency, 2010b. CORINE land cover 2006 raster data - version 13 (02/2010). URL: https://www.eea.europa.eu/ds_resolveuid/a645109f7a11d43f5d7e275d81f35c61. Acessed on 29th May 2017.

EEA, European Environment Agency, 2016. AirBase – European air quality database, version 8. URL: https://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8. Accessed on 20th April 2017.

EMEP and CEIP, 2014. Present state of emission data. URL: http://www.ceip.at/ms/ceip_home1/ceip_home/webdab_emepdatabase/reported_emissiondata/. Accessed on 29th May 2017.

Fensterer, V., Küchenhoff, H., Maier, V., Wichmann, H.E., Breitner, S., Peters, A., Gu, J., Cyrys, J., 2014. Evaluation of the impact of low emission zone and heavy traffic ban in Munich (Germany) on the reduction of PM10 in ambient air. International Journal of Environmental Research and Public Health 11, 5094–5112. doi:10.3390/ijerph110505094.

Feranec, J., Soukup, T., Hazeu, G., Jaffrain, G. (Eds.), 2016. European Landscape Dynamics: CORINE Land Cover Data. CRC Press, Boca Raton, Florida, USA.

Gilliland, F., Avol, P.K., Jerrett, M., Dvonch, T., Lurmann, F., Buckley, T., Breysse, P., Keeler, G., de Villiers, T., McConnell, R., 2005. Air pollution exposure assessment for epidemiologic studies of pregnant women and children: lessons learned from the Centers for Children's Environmental Health and Disease Prevention Research. Environmental Health Perspectives 113, 1447–1454.

Gräler, B., Pebesma, E., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. The R Journal 8, 204–218.

Haupt, H., Schnurbus, J., Semmler, W., 2018. Estimation of grouped, time-varying convergence in economic growth. Econometrics and Statistics forthcoming. doi:10.1016/j.ecosta.2017.09.001.

Haupt, H., Schnurbus, J., Tschernig, R., 2010. On nonparametric estimation of a hedonic price function. Journal of Applied Econometrics 5, 894–901. doi:10.1002/jae.1186.

Hayfield, T., Racine, J.S., 2008. Nonparametric econometrics: The np package. Journal of Statistical Software 27, 1–32. doi:10.18637/jss.v027.i05.

Hennig, F., Sugiri, D., Tzivian, L., Fuks, K., Moebus, S., Jöckel, K.H., Vienneau, D., Kuhlbusch, T.A., de Hoogh, K., Memmesheimer, M., et al., 2016. Comparison of land-use regression modeling with dispersion and chemistry transport modeling to assign air pollution concentrations within the Ruhr area. Atmosphere 7, 48.

Hijmans, R.J., 2016. raster: geographic data analysis and modeling. URL: https://CRAN.R-project.org/package=raster. R package version 2.5-8.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment 42, 7561–7578. doi:10.1016/j.atmosenv.2008.05.057.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., 2006. Spatial interpolation of ambient ozone concentrations from sparse monitoring points in Belgium. Journal of Environmental Monitoring 8, 1129–1135. doi:10.1039/b612607n.

Janssen, S., Dumont, G., Fierens, F., Mensink, C., 2008. Spatial interpolation of air pollution measurements using CORINE land cover data. Atmospheric Environment 42, 4884–4903. doi:10.1016/j.atmosenv.2008.02.043.

Li, Q., Racine, J., 2004. Cross-validated local linear nonparametric regression. Statistica Sinica 14, 485–512. doi:10.2307/24307205.

Li, Q., Racine, J., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press, Princeton, USA.

Mammen, E., Rothe, C., Schienle, M., 2012. Nonparametric regression with nonparametrically generated covariates. Annals of Statistics 40, 1132–1170. doi:10.1214/12-AOS995.

Mercer, L.D., Szpiro, A.A., Sheppard, L., Lindström, J., Adar, S.D., Allen, R.W., Avol, E.L., Oron, A.P., Larson, T., Liu, L.J.S., Kaufman, J.D., 2011. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NO_x) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Atmospheric Environment 45, 4412–4420. doi:10.1016/j.atmosenv.2011.05.043.

Montero, J.M., Fernández-Avilés, G., Mateu, J., 2015. Spatial and Spatio-Temporal Geostatistical Modeling and Kriging. John Wiley & Sons, Ltd. 266–273. doi:10.1002/9781118762387.ch8.

Nash, J.C., 2014. On best practice optimization methods in R. Journal of Statistical Software 60, 1–14. doi:10.18637/jss.v060.i02.

Nash, J.C., Varadhan, R., 2011. Unifying optimization algorithms to aid software system users: optimx for R. Journal of Statistical Software 43, 1–14. doi:10.18637/jss.v043.i09.

Pagan, A., 1984. Econometric issues in the analysis of regressions with generated regressors. International Economic Review 25, 221–247. doi:10.2307/2648877.

Pebesma, E.J., 2004. Multivariable geostatistics in S: The gstat package. Computers and Geosciences 30, 683–691. doi:10.1016/j.cageo.2004.03.012.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, AT. URL: http://www.R-project.org/.

Rmetrics Core Team, Wuertz, D., Setz, T., Chalabi, Y., Maechler, M., Byers, J.W., 2015. timeDate: Rmetrics – chronological and calendar objects. URL: https://CRAN.R-project.org/package=timeDate. R package version 3012.100.

Robinson, D., 2017. broom: convert statistical analysis objects into tidy data frames. URL: https://CRAN.R-project.org/package=broom. R package version 0.4.2.

Ryan, P.H., LeMasters, G.K., 2007. A review of land-use regression models for characterizing intraurban air pollution exposure. Inhalation Toxicology 19, 127–133. doi:10.1080/08958370701495998.

Sahsuvaroglu, T., Arain, A., Kanaroglou, P., Finkelstein, N., Newbold, B., Jerrett, M., Beckerman, B., Brook, J., Finkelstein, M., Gilbert, N.L., 2006. A land use regression model for predicting ambient concentrations of nitrogen dioxide in Hamilton, Ontario, Canada. Journal of the Air & Waste Management Association 56, 1059–1069. doi:10.1080/10473289.2006.10464542.

Sperlich, S., 2009. A note on non-parametric estimation with predicted variables. Econometrics Journal 12, 382–395. doi:10.1111/j.1368-423x.2009.00291.x.

Wang, R., Henderson, S.B., Sbihi, H., Allen, R.W., Brauer, M., 2013. Temporal stability of land use regression models for traffic-related air pollution. Atmospheric Environment 64, 312–319. doi:10.1016/j.atmosenv.2012.09.056.

Wolf, K., Cyrys, J., Harciníková, T., Gu, J., Kusch, T., Hampel, R., Schneider, A., Peters, A., 2017. Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany. Science of the Total Environment 579, 1531–1540. doi:10.1016/j.scitotenv.2016.11.160.

3.A Appendix

3.A.1 Tables and Figures

Table 3.A.1: Optimised class weights. Following Janssen et al. (2008), class weights a_2, a_{10} and a_{11} are set to 1, 0 and 0, respectively. Therefore the optimisation procedure returns optimal values for the other eight class weights.

Germany	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
QL	2.77	1.00	0.92	0.73	0.71	0.09	0.34	0.10	0.36	0.00	0.00
LL	2.96	1.00	0.92	0.80	0.67	0.08	0.31	0.10	0.35	0.00	0.00
TypeQL	1.76	1.00	1.39	2.09	1.52	1.47	0.91	0.12	0.13	0.00	0.00
TypeLL	3.53	1.00	1.77	3.65	2.21	2.31	1.25	0.33	0.47	0.00	0.00
NP	0.05	1.00	2.07	5.07	1.16	1.17	4.25	2.21	0.81	0.00	0.00
Belgium											
QL	3.49	1.00	1.49	6.00	2.75	1.38	1.73	0.35	0.00	0.00	0.00
LL	1.62	1.00	1.63	3.65	2.10	1.30	1.80	0.40	0.00	0.00	0.00
TypeQL	0.83	1.00	0.96	2.42	1.65	0.95	1.11	0.27	0.00	0.00	0.00
TypeLL	0.89	1.00	1.09	3.16	1.91	0.91	0.13	0.36	0.00	0.00	0.00
NP	0.98	1.00	2.61	6.02	1.10	1.12	3.78	0.75	0.63	0.00	0.00

* with first and second largest LUC ** with first, second and third largest LUC

Table 3.A.2: Results of LOOCV for different specifications and their predictive performance with regard to RMSE.

Germany	QL	LL	TypeQL	TypeLL	NP	NPnoBeta*	NPnoBeta**
Background	16.70	16.70	12.79	12.93	14.18	13.16	13.18
Industrial	14.52	14.46	13.10	13.25	14.25	15.29	15.45
Traffic	27.06	27.02	25.30	25.63	25.98	25.52	25.16
Overall	20.84	20.82	17.99	18.21	19.07	18.43	18.31
Belgium							
Background	13.16	13.02	12.88	12.80	13.40	13.33	13.57
Industrial	14.63	14.92	14.33	14.09	13.99	14.86	15.10
Traffic	14.02	14.04	29.19	14.81	13.84	13.69	14.35
Overall	13.76	13.79	15.69	13.51	13.66	13.88	14.19

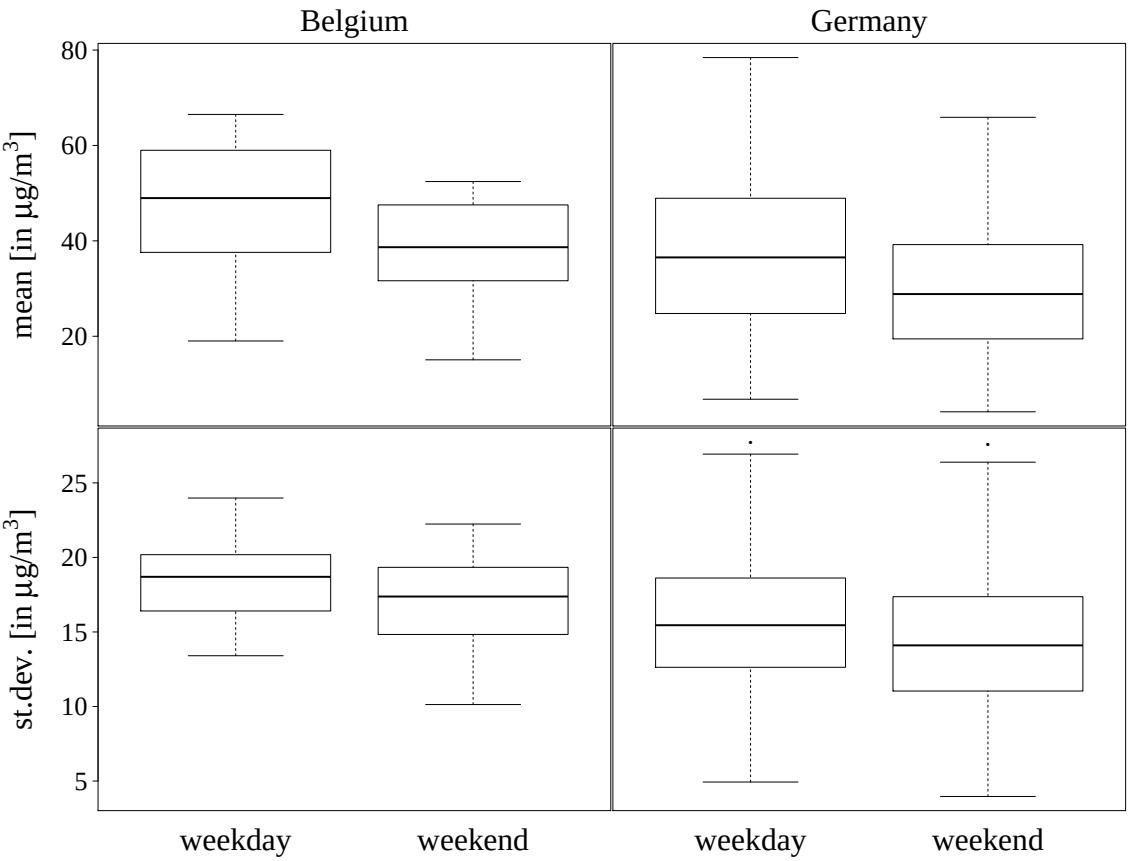

Figure 3.A.1: Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian background site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

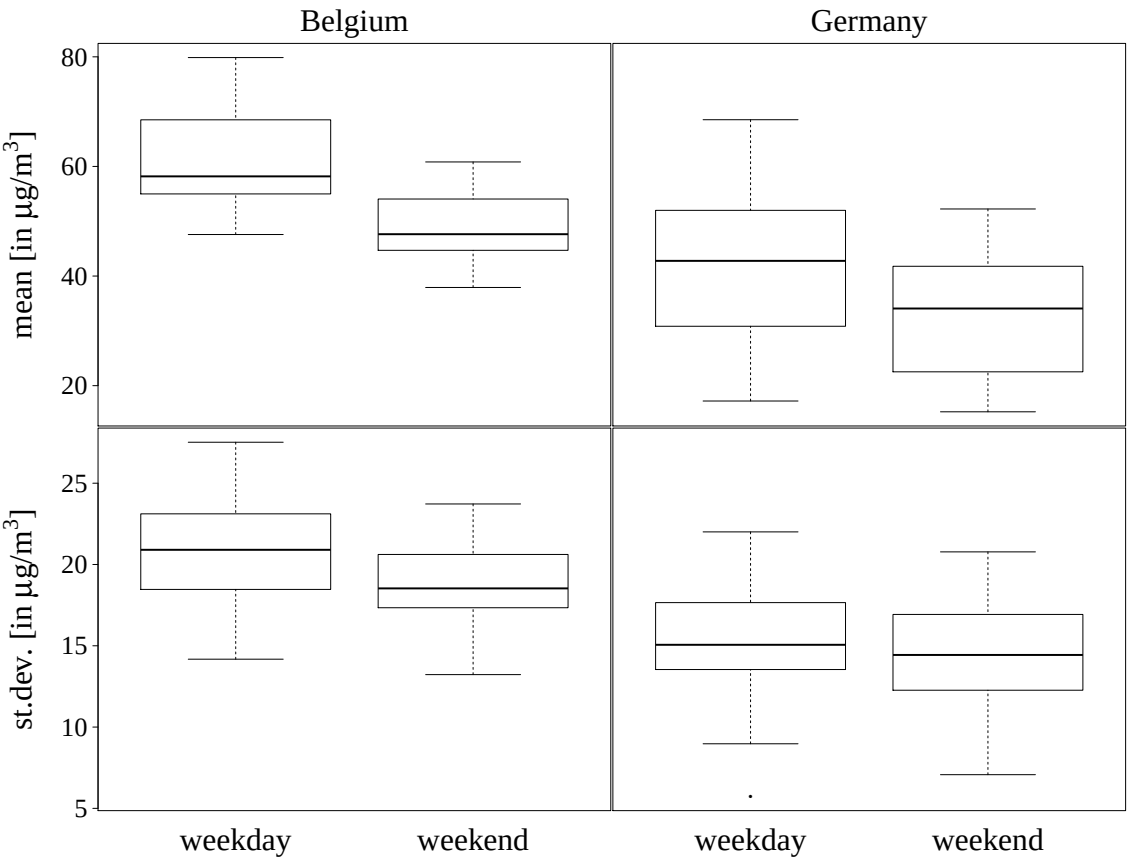

Figure 3.A.2: Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian industrial site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

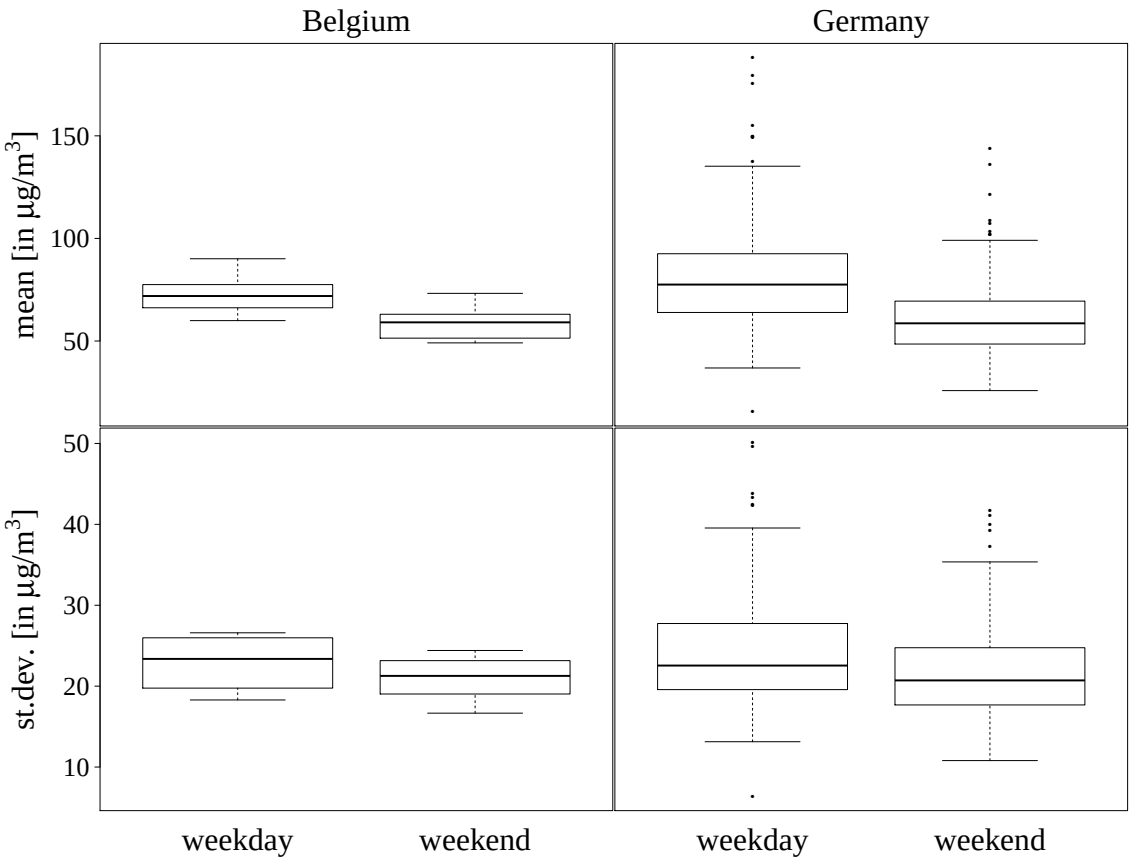

Figure 3.A.3: Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian traffic site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

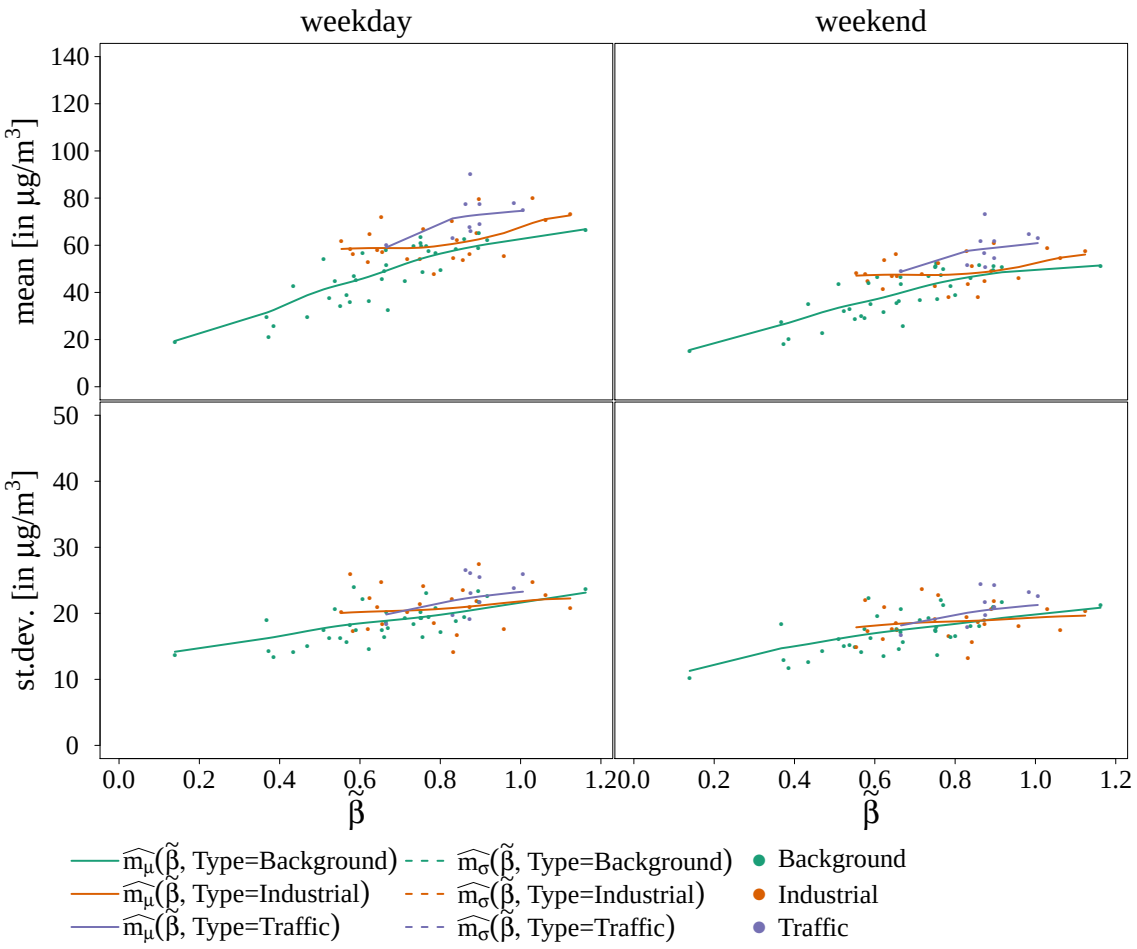

Figure 3.A.4: Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to the nonparametric approach (specification NP).

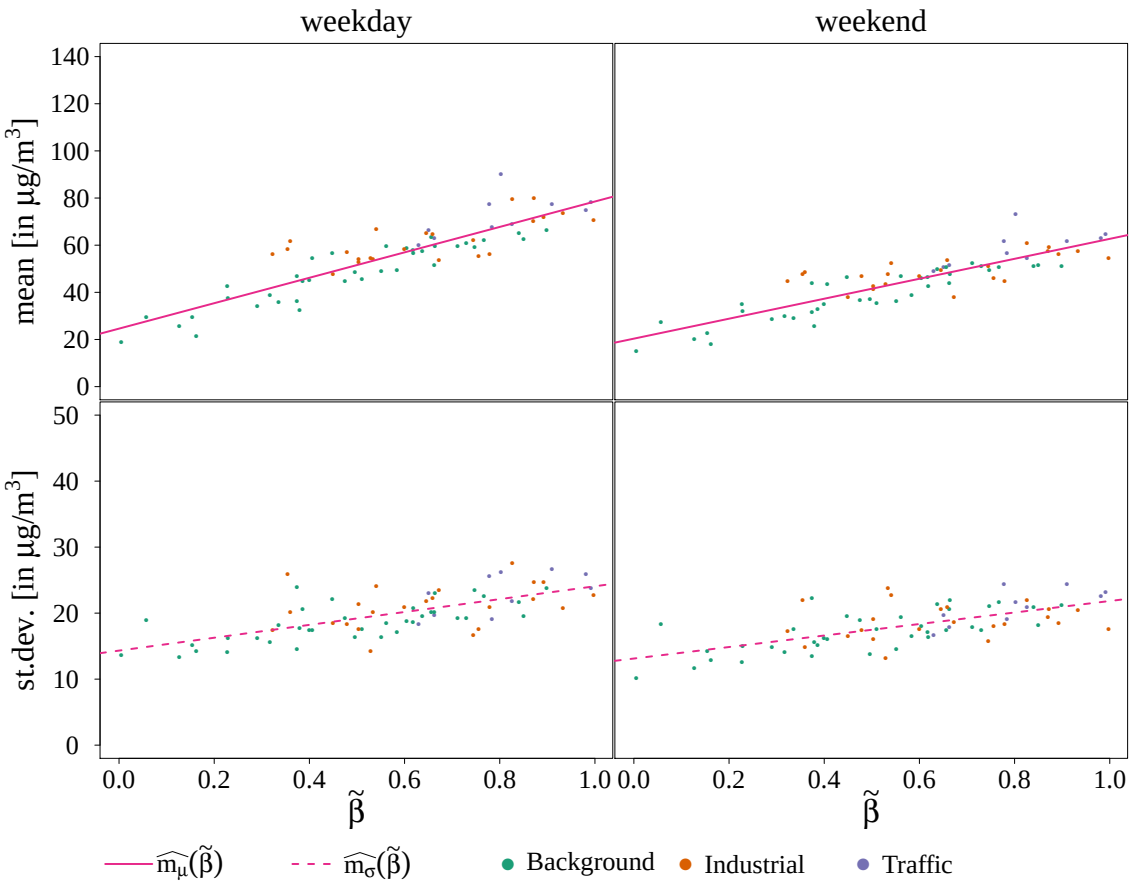

Figure 3.A.5: Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation (specification LL).

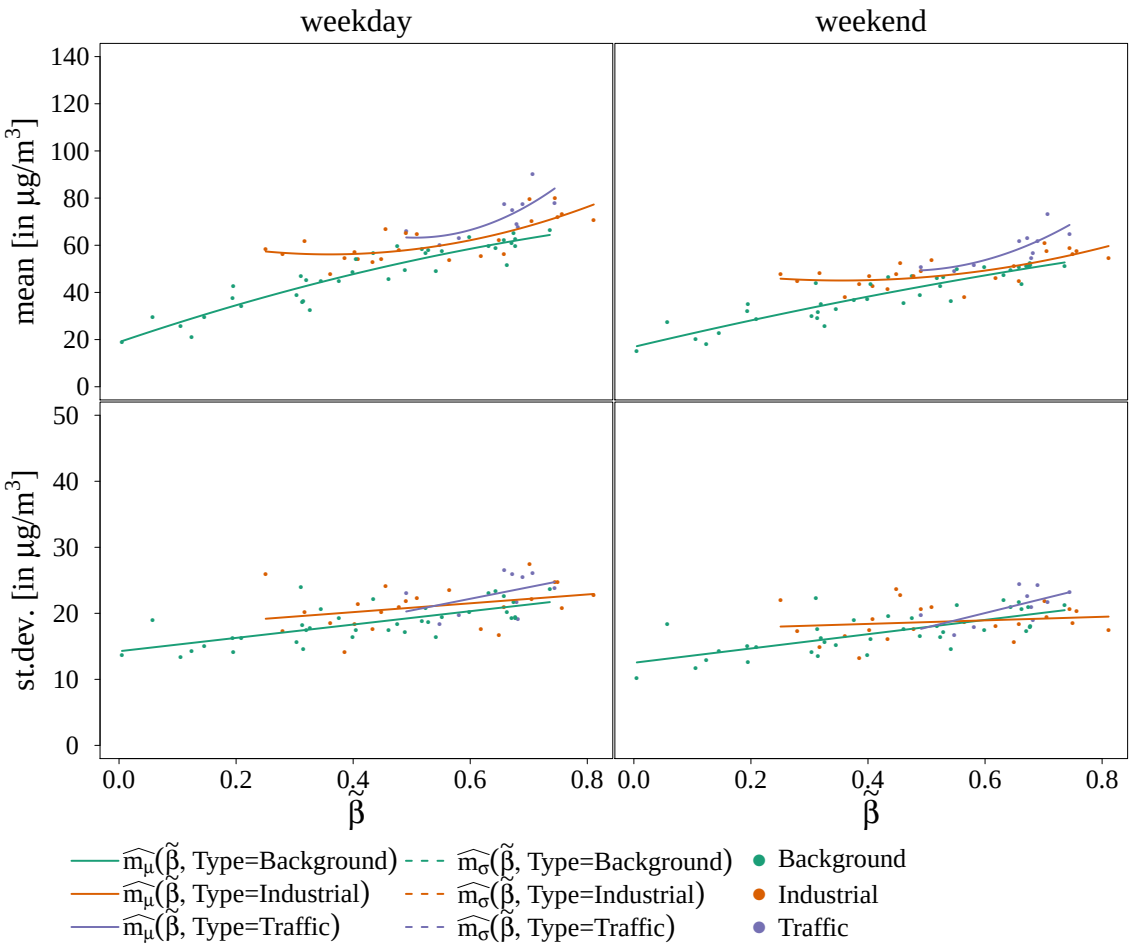

Figure 3.A.6: Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeQL).

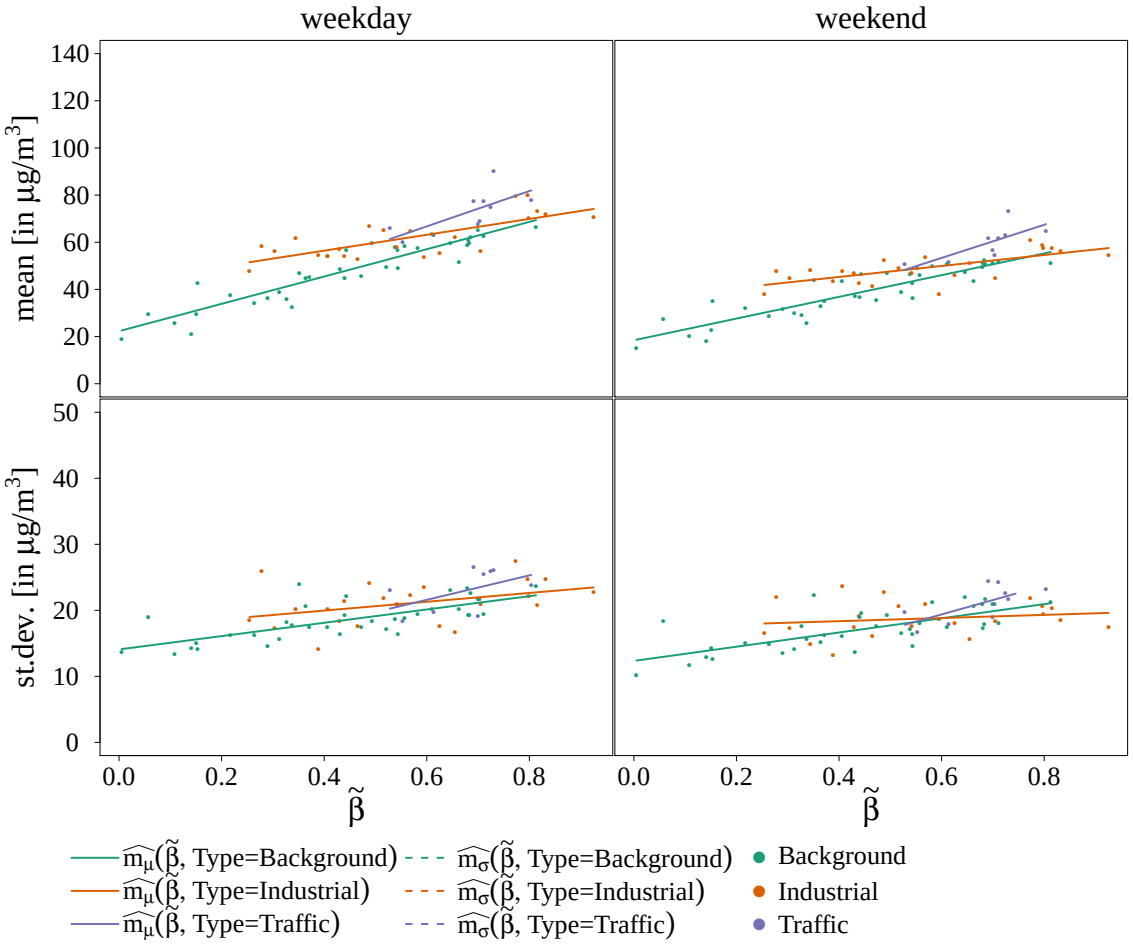

Figure 3.A.7: Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeLL).

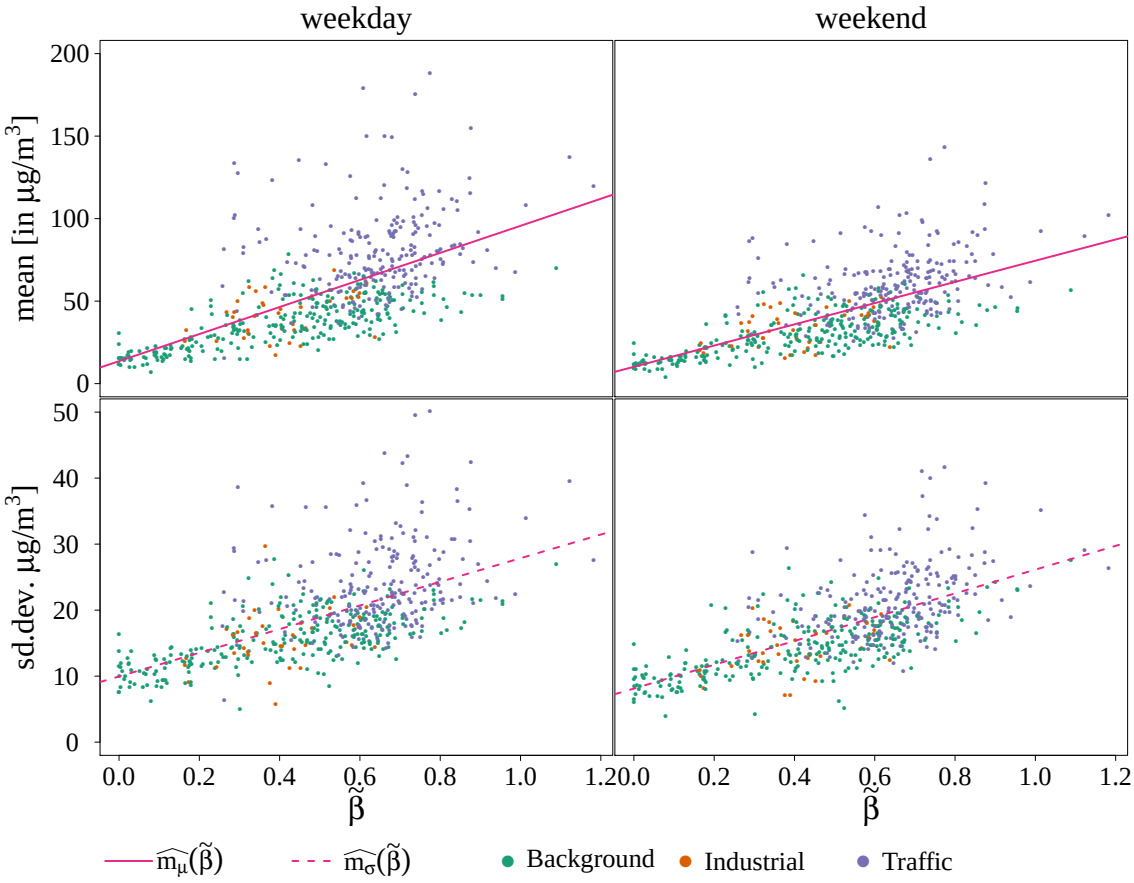

Figure 3.A.8: German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation (specification LL).

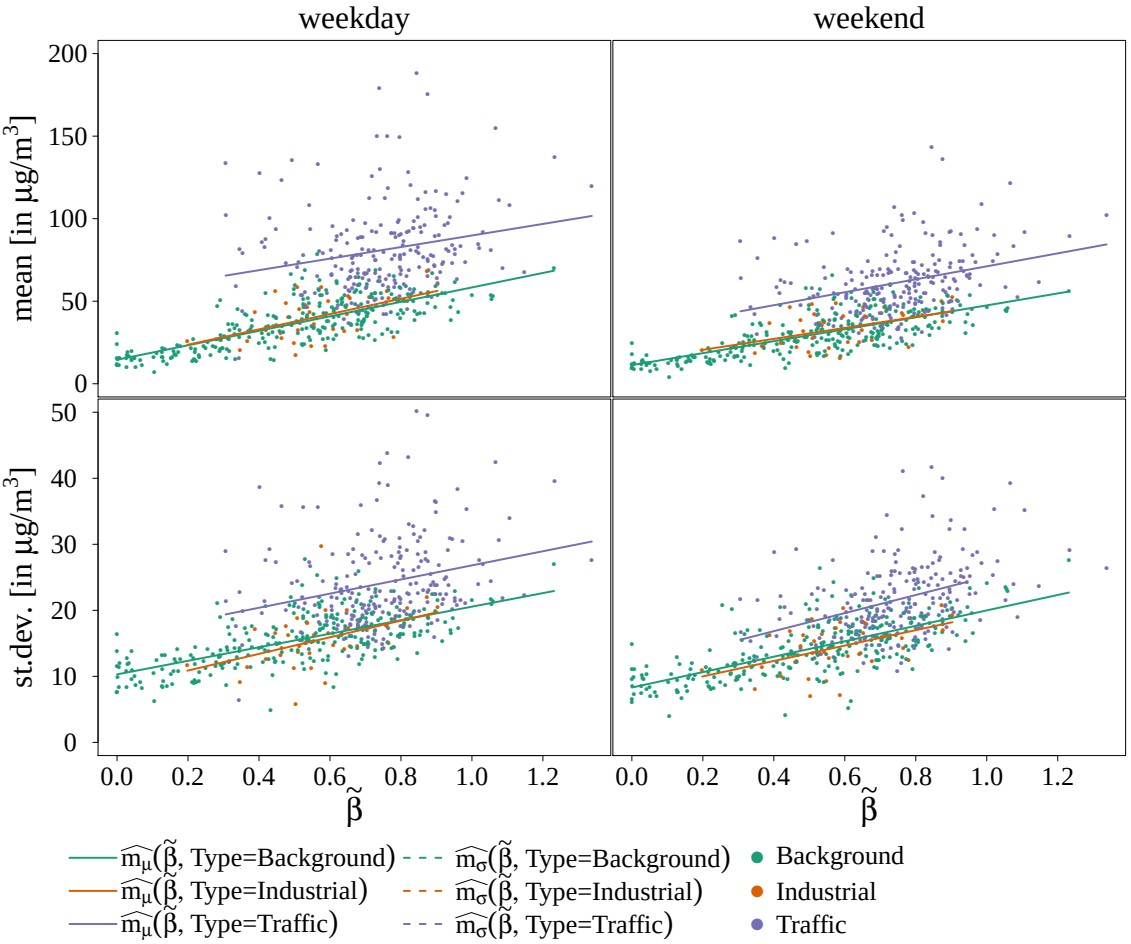

Figure 3.A.9: German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeLL).

3.A.2 Data related descriptions

Metadata in AirBase

AirBase consists of monitoring data from fixed monitoring sites as well as meta-information on the monitoring sites involved. The following meta-information is provided by AirBase: station european code, station local code, country iso code, country name, station name, station start date, station end date, type of station, station ozone classification, station type of area, station subcat rural back, street type, station longitude deg, station latitude deg, station altitude, station city, lau level1 code, lau level2 code, lau level2 name, EMEP station.

With regard to air pollution analysis the following variables might be of interest:

- type of station - Background, Industrial, Traffic
- station ozone classification - rural, rural background, suburban, urban (the pollutants NO_2 and O_3 are strongly correlated, see Janssen et al., 2008, p. 4889)
- station type of area - rural, suburban, urban
- station subcat rural back - near city, regional, remote
- street type - Canyon street (L/H $<$ 1.5), Highway (average speed vehicles $>$ 80km/h), Unknown, Wide street (L/H $>$ 1.5); length (L) of the canyon usually expresses the road distance between two major intersections; height (H) of the canyon
- station longitude deg
- station latitude deg
- station altitude

In our work we consider station longitude deg, station latitude deg and type of station.

Data processing and data quality AirBase

In the following we describe how we have processed the hourly recorded NO_2 values and provide information about the data quality. Quality flags in the raw data of the AirBase statistics indicate the quality of each measurement value. A quality flag > 0 indicates valid measurement data. A quality flag $<= 0$ indicates invalid or missing data (EEA, European Environment Agency, 2016).

Belgian AirBase data: The time period 1st Jan 2001 to 31st Dec 2006 has $24 * (365 * 6 + 1) = 52\,584$ hours. A full sample with recorded hourly values for each of the 70 monitoring sites would therefore consist of $52\,584 * 70 = 3\,680\,880$ observations. There is no entry in the source data for 815 064 site-date-hour combinations, which corresponds to about 22.14%. This is partly due to the fact that some sites have not recorded the NO_2 concentrations over the whole period, either they have been built up after 1st Jan 2001 or switched off before 31st Dec

2006 or for some time between the 1st Jan 2001 and the 31st Dec 2006. The percentage of either missing or not validated entries in the source data is equal to $371\,497/(3\,680\,880 - 815\,064) \widehat{=} 13.43\%$. We have omitted missing and non validated values from further analysis and have extracted from the daily maximum NO_2 concentration for each site-day combination the remaining data which results in 112 340 maximum values, compared to $70*(365*6+1) = 153\,370$ maximum values if data for each site-date combination existed. The Belgian data do not contain any extremely high values (above 500 $\mu g/m^3$) nor any negative daily maximum values.

German AirBase data: The time period 1st Jan 2007 to 31st Dec 2012 has $24 * (365 * 6 + 2) = 52\,608$ hours. A full sample with recorded hourly values for each of the 537 monitoring sites would therefore consist of $52\,608 * 537 = 28\,250\,496$ observations. There is no entry in the source data for 5 391 528 site-date-hour combinations, which corresponds to about 19.08%. This is partly due to the fact that some sites have not recorded the NO_2 concentrations over the complete time, either they have been built up after 1st Jan 2001 or switched off before 31st Dec 2006 or for some time between the 1st Jan 2001 and the 31st Dec 2006. The percentage of either missing or not validated entries in the source data is equal to $1\,547\,472/(28\,250\,496 - 5\,391\,528) \widehat{=} 6.77\%$. We have omitted missing and non validated values from further analysis and have extracted from the daily maximum NO_2 concentration for each site-day combination the remaining data which results in 920 343 maximum values, compared to $537 * (365 * 6 + 2) = 1\,177\,104$ maximum values if data for each site-date combination existed. Omitting missing and non validated values reduces the number of sites from 537 to 536. Further investigation has shown that the source data do not contain any validated data for site `DETH082`. Three daily maximum values have been removed as they are extremely high (above 500 $\mu g/m^3$) and 58 as they are negative such that finally 920 282 maximum values and 536 sites remain for further analysis.

Chapter 4

The Role of Blueprints in Quantitative Cultural Comparison

In coauthorship with Dr. Jörg Scheffer, Faculty of Arts and Humanity, Geography Section, University of Passau

Chapter Abstract

Researchers in the field of Comparative Cultural Analysis have adopted tools for exploring the ever-growing data sets from social networks and other sources into their common set of methods – a development often called the "computational turn" in social sciences. More traditional approaches to comparative analysis are based on thorough sampling, model development, and extensive testing. Strong statistical assumptions on model homogeneity for the measurement of beliefs, attitudes, and other cultural phenomena are needed for these models, in order to achieve a maximum degree of objectivity. Especially the sociological perspective on culture relies heavily on ready-made blueprints for the identification of culturally homogeneous groups. We ask what lessons can still be learned from these "old" models and vertices when using modern techniques in comparative social studies. We discuss the advantages and drawbacks of strong model assumptions and blueprint-like repetition of cultural association. While a "big data" perspective usually uses simple comparison, a "small data" perspective can direct researchers' attention to potential problems with sampling and validity of results. We illustrate our thoughts with calculations based on data from the European Values Study. Thereby, we contribute to the critical discussion of data analysis methods in the context of cultural analysis.

4.1 Introduction

Despite the ever-growing connectivity of the world, cultural differences between people are far from vanishing. Practitioners keep underlining the importance of developing a sensibility for subtle cultural differences in different parts of the world, and international business literature warns about potential pitfalls in cross-cultural exchange (e.g. Moran et al., 2014). Within the last years, the availability of ever more and qualitatively new data sources have given fresh impetus to research on cultural differences in many scientific disciplines. Video, image, and text data provide new evidence and material for investigation. While theoretically, there is no need to group people when there is so much granularity and detail available, methodological practice shows that despite globalization, cultural blueprints that set culture equal to nationality have lost nothing of their appeal for comparative cultural analysis (Manovich, 2018).

The turn to computational methods for cultural analysis, however, puts the differentiated, critical perspective of cultural researchers on quantitative analysis in peril. As Masson et al. (2017, p.26) point out, some researchers in humanities fear that "[b]y succumbing to the lure of scientism [...] humanists run the risk of forgetting what they excel at – critical interpretation – and by the same token, of impoverishing their practice". Consequently, the question of whether embracing big data analytics is beneficial to social sciences or not is controversial. It is a fact, however, that over the years, data-centric research in the humanities evolved from a supporting tool to a research area of its own – *Digital Humanities* (Masson et al., 2017).

With this article, we investigate the role of small-data models based on strict measurement models and sociological, group-based blueprints for the mediation of cultural differences. We depart from the idea that it is possible to apply a context-specific perspective to cultural differentiation and provide an introduction to the underlying intuition to comparative, quantitative cultural research. Abstract conceptualizations of specific cultural elements like convictions must, on the one hand, be very concise and distinguishable from other, related concepts. Measurement scales need to be objective and non-ambivalent when comparing cultures (Poortinga,

1989). On the other hand, they have to be complex enough to evoke real interest and allow for an actual gain of knowledge. Structural equation models allow for verifying the applicability and comparability of fairly complex concepts between different groups, as well as a modeling of their interplay and statistical inference (compare Kirby and Bollen, 2009).

The main goal of this article is to clarify the role of group-based boundaries in quantitative analysis of unobserved cultural patterns. We seek to answer the question of whether it is possible to combine an objective perspective on measurement models with a differentiated concept of cultural identity and group membership. We use data from the *European Values Study* (EVS) (GESIS Data Archive for the Social Sciences, 2018) to empirically illustrate what a regionally and contextually differentiated comprehension of culture can look like. The basic idea behind our approach is that there is no need for a globally unique system of classification when the goal of analysis lies in the description of contextually relevant differences. The conceptualization of space should much more take into consideration the identification of characteristics that dominate a specific situation of interaction in a particular setting.

We provide a general introduction to the methodology of comparative cultural research, explain the challenges associated with standard concepts, and discuss how cultural and physical space can be reconciled (Section 4.2). Building upon this framework, we discuss to what degree quantitative methodology can meet the requirements of a differentiated but manageable operationalization of culture and space (Section 4.3). Our methodological explanations cover structural equation modeling, group-based measurement model validation, and hierarchical models. Two case studies for measurement of culture rooted in the fields of institutional religiosity and solidarity accompany our considerations for empirical illustration. We discuss the main insights of our study in Section 4.4. Section 4.5 concludes.

Our contribution is as follows. We provide an intuitive introduction of social scientists to the concepts of comparative cultural research with a focus on the role of "groups" in statistical models. The article concentrates on the question to what extent it is possible to quantitatively assess cultural differences and, at the same

time, avoid the traps of stereotype boxing of culture. We show that the close association of cultural space with political borders is necessary to provide explicit references for comparison, but that instead of the ever-repeated national blueprints, more differentiated, context-specific conceptions of cultural regions are needed.

4.2 Are Cultural Blueprints Necessary?

Before going into the details of small-world models for cultural comparison, we will outline the Computational Turn in social sciences and motivate our research by pointing out the risks that arise with the usage of large, non-curated data sets for cultural research. Also, we will specify the Research Questions we aim to answer with the ensuing analysis and review.

4.2.1 The Computational Turn in Social Sciences

Over the last decades, digital technology and progress in methodology for the analysis of large data sets has fundamentally changed the research in social sciences. Digital technology has become a mediator of research. The field of *digital humanities* puts the digital creation of culture into the center of its interest – to analyze how technological mediation affects the discipline of humanities itself (Berry, 2011). Ontologies and epistemic foundations change, also *within* data-enhanced comparative cultural research. New data sources are available in new magnitudes of size. Models for highly controlled and curated datasets are now applied to more massive, uncontrolled data sets (Miller, 2010). However, this carries risks.

> [S]caling small data exposes them to the new epistemologies of data science and to incorporation within new multi-billion data markets being developed by data brokers, thus potentially enrolling them in pernicious practices such as dataveillance, social sorting, control creep, and anticipatory governance, for which they were never intended (Kitchin and P. Lauriault, 2015, p.464).

Consequently, the risks of concluding from large datasets from social media or other sources require some attention. Models for the analysis of curated, high-quality

survey data rely on the high degree of scarcity, stability, and cleanliness in such contexts. They depend on statistical assumptions such as independence, stationarity, and normality. Moreover, ideally, the theory that should be validated with the data set *precedented* the analysis (Miller, 2010).

Old habits are still in place. It is, for example, a common practice to use these data sets to compare different nationalities:

> Given that certain demographic categories have become taken for granted in our thinking about society, it appears natural today to group people into these categories and compare them in relation to social, economic, or cultural indicators (Manovich, 2018, p.8).

When using traditional models and algorithms in a new different context, the underlying epistemic and statistical assumptions have to be thoroughly understood. The field of *critical data studies* put the "thinking critically about and research algorithms"(Kitchin, 2017, p.16) at the heart of its practice. Algorithms can produce evidence that leads to contingent knowledge – contingent on its current socio-technical context. In a similar vein, we investigate the power and politics of confirmatory multi-group factor analysis for cultural comparison and what kind of conclusions can be drawn from such analysis, considering its framing within a broader socio-technical context of comparative cultural analytics. With this approach, we follow the recommendation for critical data studies by Barocas et al. (2013) and Kitchin (2017).

With our perspective on comparative cultural analytics, we engage in a *sociological* perspective on culture, where common characteristics, attitudes, or behaviors are attributed to pre-defined groups. The tradition of positivist social science, looking for patterns in social behavior, heavily relies on the statistical study of the correlation between measurable characteristics of humans, distinguishing between *dependent* and *independent* variables (compare Manovich, 2018). The mapping of culture to social groups is a simplification and prerequisite for comparison. It will be part of our consideration of how this simplification can be mitigated in analysis when societies imply "several levels of complexity" (Conte et al., 2012, p.334), and culture does neither describe an overly detailed individual, nor an overly generalized national

lever, but also an intermediate levels of communities.

4.2.2 The Persistence of Political Blueprints for Cultural Comparison

Globalization has not wiped out deep-level cultural differences between societies. Rather, the increase in the number of intercultural interactions augments the importance of knowledge about differences to avoid misunderstandings, tensions, and conflicts Ang and Van Dyne (2015); Poortinga (1989); Delanty (2003); Everitt (2006). Inter-cultural conflicts comprise debates on the integration of refugees (Thran and Boehnke, 2015), and contemporary surges of nationalism(compare Anderson and Keil, 2016) speak in favor of a growing emphasis on the national character of culture.

Sighting practitioners' literature reveals that the national reference has found its way into both quantitative and qualitative research. Guidelines on intercultural communications and training programs regularly fall back on the well-known pioneering work by Geert Hofstede (Hofstede, 2003). Hofstede provided a system of indexes on the so-called *dimensions* of cultures. These indexes boil practically relevant parameters of behavior down to a dichotomous scale, on which every single nation can be positioned from "low" to "high". Examples include the handling of uncertainty, the perception of the role of the individual, and the centrality of power. The scales provide a simple frame of reference in daily contact with "foreigners", leading to direct and easily understandable consequences in practical interaction (Hofstede, 2001, 2003).

Quantitative comparison of nations has repeatedly proved to be useful for the explanation of misunderstandings in a bi- or trilateral context. Cross-cultural analysis in the "spirit of Hofstede" does not claim to be valid for every individual, but generally describes dominant tendencies within a nation. It aims to provide a framework of orientation and insights into the dimensions where problems *might* show up. Hofstede (2002) underlines that his method concentrates on single aspects (dimensions) of culture, and shows no ambition to paint the complete and detailed picture of every possible cultural constellation. However, the comparative yardstick is still inflexibly linked to a national assignment, and bases on a highly schematized gen-

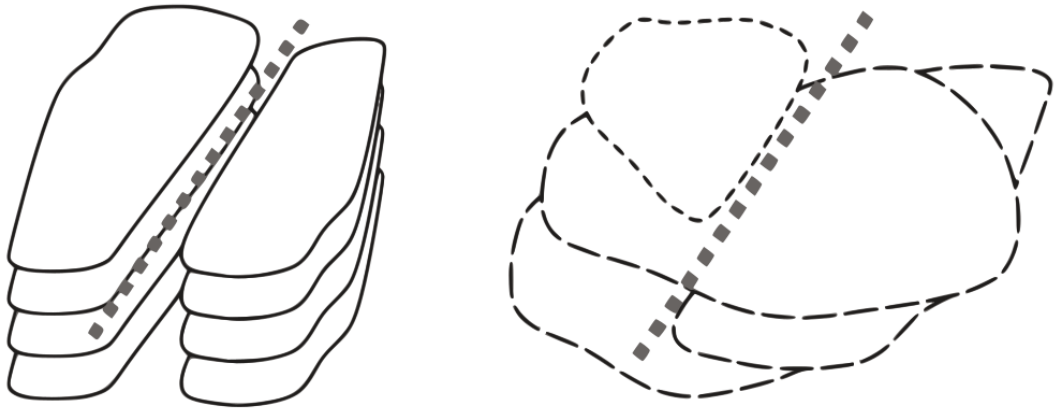

Figure 4.1: Different concepts of cultural comparison. The dotted line represents a political (national) border. The left panel represents nationality-based cultural comparison, while the right panel shows a selective perspective on single cultural layers.

eralization of space and measurement models (McSweeney, 2002). As dimensions can only exist within the static boxes of pre-fabricated concepts, every comparison automatically leads to the reproduction of the units. This way, over time, cultural contrasts become an artifact of the method instead of a good approximation to reality. Even though different boxing of physical space might lead to more relevant or more adequate findings, the concept of the congruence of political and cultural borders perpetuates.

To avoid this deterministic perspective on culture, and in order to more flexibly account for the needs of specific research interest, models need to take both geophysical embedding and distance of survey participants into account. An alternative to national grouping requires finding a different way to link people to cultural communities, a way that allows for flexible adaptations to regionally dominant concepts. There is scope for a new conceptualization of comparative methodology. The spatial embedding of a population and its cultural beliefs and attitudes allows uncoupling cultural attribution from national identity. Every cultural concept can have a distinctive form of representation. Such a categorization of culture is validated only within a specific research question and represents one of many possible perspectives on culture. The left panel of figure 4.1 represents group-based cultural comparison. As comparison is based on always the same political blueprints, whatever the cul-

tural construct of interest, the analysis ignores any cross-border cultural similarity. Vice versa, always using the same spatial blueprint leads to the sorting out of partially interesting aspects that do not show significant differences between the groups compared. The right panel represents the concept of *selective cultural comparison*, which does not compare *different cultures*, but *collectives with different cultural characteristics*, represented by the single layers. These can be independently described in individual spatial shapes, allowing for the adaptation of the analytic framework to a specific interest. It is thus possible to reveal both cross-border similarities and differences.

In the discussion of the adequate form to frame cultural differences, there are wide disparities regarding the need for pre-defined groups, i.e., blueprints. Some researchers emphasize the risk of inappropriate homogenization within spatial borders (e.g. McSweeney, 2002). As Rippl and Seipel (2008, p. 18) point out, most studies that claim to compare cultural collectivity could more adequately be labeled cross-country comparative studies. The undifferentiated usage of the notions "country", "society" and "culture" has been criticized repeatedly, and many authors have demanded a joint effort for the deconstruction of such thinking patterns (Hermans and Kempen, 1998; Straub, 2003). Nevertheless, the pragmatic ease of clear-cut spatial delineation still dominates the discussion of intercultural differences (see also Triandis, 1994, pp. 8-9).

This pragmatic choice is not without reason. Any mediation of cultural differences is preceded by the conceptualization of culture itself, and by the question of how cultural groups can be differentiated. While nationality is not the only source of cultural identity, "the other" is rhetorically still closely associated with nationality. Cultural collectivity arises in the interplay of different sources of enculturation: local and global sources of culture and identity comprise but are not limited to nationality. However, trying to assess every aspect of culture inhibits the denomination of a concrete object for comparison. The evaluation of cultural collectivity requires a well-defined, tangible object and a comparative standard or "yardstick". Both must be valid within every group that is to be compared to each other, to avoid comparing apples with oranges. Objective criteria of equivalence need to be verified for any

measurement models and for the test procedure itself (compare, e.g. Van Deth, 1998). The total abstention from spatial boundaries and classification of territories inhibitis the description of collectivity, and questions the existence of "a culture" in the sense of a characterization of a group of people, which can be compared to other cultures. Grouping is needed to give a body to the collective one is describing; and as nationality is an easily digestible concept for grouping, many accept the national labels as a pragmatic approach that serves for the transmission of the message of cultural differences to an easily digestible level (e.g. Antonczyk and Salzmann, 2014; Ahern et al., 2015).

It is unsure whether the required objectivity is realistic after all. The defendants of a culturally intrinsic (i.e. *emic*, culturally immanent) conceptualization of culture depart from the idea that schemes of thinking and acting can only be understood within their original systemic context. Consequently, there are only few particular similarities with any other cultures. A common way to circumvent the total denial of comparison and to find comparable concepts in a qualitative approach is to analyze situations of conflict between members of two different cultures. In the collaborative analysis of the situation, so-called *cultural standards* manifest themselves and can be described qualitatively as contrasting pairs (Tomas, 1996). Nationality then serves to label the total of differences between two persons, which assumes external validity of the findings for peer groups. Quantitative comparison results from a cross-culturally overarching, *etic* perspective, i.e., it assumes that it is possible to have a neutral measurement instrument that provides accurate measures. An *etic* perspective assumes that it is possible for the researcher be neutral: she can anlyzes culture without being influenced by his or her own culture. The concepts which are to be compared are gained from an outside perspective. The operationalization of measurement models via standardized surveys creates the basis for a massive collection of data to measure the concepts (Helfrich, 2003). Collective patterns summarize statistically equivalent answers. Again, political pre-defined borders are helpful for the process of comparison - the clear-cut grouping is the basis for verification of the validity of constructs across groups.

National blueprints risk to reduce "the Italian" and "the German" to their in-group

similarity and between-group differences. They inhibit a differentiated consideration of local specificity and cross-border similarity. Despite its aim to make cultural differences comprehensible, cross-cultural comparison practice may have contributed to this situation. Some researchers underline that the pragmatic, infinite reproduction of readily available national classifications can lead to the transformation of political boundaries into cultural clichés (e.g. McSweeney, 2002). Nevertheless, the equalization of culture and nation is present in many studies with an inter-cultural disclaimer. National blueprints provide clear-cut references and obvious cultural group reference. Therefore, they are in regular use in inter-cultural research in all kinds of scientific fields, like labor economics (Tubadji et al., 2014), mergers and acquisition (Ahern et al., 2015), financing (Antonczyk and Salzmann, 2014), marketing, and inter-firm cooperation (Gesteland, 2002), and many more. The simple categorization as nations that incorporate one specific culture allows for structuring collectivity and to keep it concise and empirically comparable. At the same time, there are several layers of culture, which leads to social complexity. Societies do not consist of individuals embedded in a homogeneous nationality - instead, groups, tribes, networks, communities, and other peer groups put several layers between the micro and macro level.

Moreover, these layers interact. *How* they interact so is the object of vivid research (Conte et al., 2012). Within Europe, there are not only groups within groups, but there is also important between-group exchange, caused, for example, by flows of migration (compare Tomas, 1996). This does not only lead to the mixing of distinct cultures, but also the emergence of new hybrid or transcultural imprints on an individual level (Featherstone, 1995; Kraidy, 2005; Martin and Nakayama, 2010). It is not trivial to find a proper equilibrium between an overly simplifying national categorization of people and an overly complicated deconstruction of homogeneous cultural collectives.

4.2.3 Research Questions

In light of the need for methodological guidelines in quantitative cultural research, we can formulate the following research questions.

Research question 1 (RQ1): How can cultural collectivity be quantitatively assessed and compared across different regions?

We will outline the role of fixed political or spatial units for the discussion of cultural differences, translating short-term cultural phenomena like attitudes and opinions and long-term cultural patterns like ethics and values into statistical models. Focusing on cultural *values*, we explain two measurement models concerning *church adequacy*, a part of the concept of *institutional church religiosity* and *social solidarity* as a more specific concept underlying general concepts of *solidarity*. We will discuss to what extent the *a priori* definition of clear-cut cultural regions is necessary for cross-cultural research. An empirical discussion with a concrete example accompany our discussion, where we assess the in-group and between-group homogeneity of concepts that express culture by mapping them on different local scales.

Research question 2 (RQ2): Which statistical methods and models can ensure comparability of different measurements?

This research question addresses the question of whether it is possible to forgo spatial blueprints and grouping when seeking to model culture. Answering research question 1 requires the application of a measurement model across several countries. For answering question 2, it is necessary to dive into the methodology of validation of measurement models. We provide an introduction to group-based structural equation modeling and show that the formation of some prior groups is necessary to obtain objective, comparable yardsticks for cross-cultural comparison. Nevertheless, *hierarchical* model specification overlays regional and supra-regional effects and accounts for the close relationship between "neighboring" regions. Finally, we provide an outlook to the exploratory methods that investigate geotagged social network data for their ability to assess "local culture".

Research question 3 (RQ3): How can the consequences of an endless repeating of national blueprints be translated into statistical terms?

The ever-repeated usage of political borders leads to the preselection of measurement models that have shown significant political patterns *before*, and to the negligence of models that did not do so. When subsequent studies, like ours, base upon these

models, they use implicit knowledge on likely results of the analysis. We will show that this thought is implicitly closely connected to the *Bayesian* school of statistical thinking and leads to problems with standard statistical tests on parameter values.

Research question 4 (RQ4): How can the model be specified with more flexibility concerning grouping, and allow for cultural "neighborhood"? While we do assume that physical space plays a vital role in the assessment of culture, we do not believe that this role is bound to the national blueprints. Assuming that cultural grouping in nationality is losing in importance, how can this be integrated into a model? How can spatial cultural "neighborhood" in a measurement model be taken into account? This research question sheds some light on spatial dependency and exchange processes.

4.3 Statistical Modeling of Cultural Collectivity

In the upcoming sections, we discuss the empirical implementation of different statistical models for cultural comparison, which we accompany by an implementation in `R` (`R` Core Team, 2013). The most important packages used for our empirical application are `lavaan` (Rosseel, 2012), `psych` (Revelle, 2014), `semPlot` (Epskamp, 2014), `semTools` (`semTools` Contributors, 2016) and `sp` (Bivand et al., 2013). Moreover, we use `classInt` (Bivand, 2017), `ggmap` (Kahle and Wickham, 2013), `GISTools` (Brunsdon and Chen, 2014), `GPArotation` (Bernaards and Jennrich, 2005), `gstat` (Benedikt Gräler and Heuvelink, 2016), `Hmisc` (Harrell Jr, 2017), `latticeExtra` (Sarkar and Andrews, 2016), `RColorBrewer` (Neuwirth, 2014), and `rgdal` (Bivand et al., 2017).

4.3.1 RQ1: Structural Equation Modeling and Group Means

In recent years, extensive data that implicitly reflect cultural opinions, attitudes, and patterns became available through the usage of social networks (for the analysis of social networks, compare Snijders, 2001; Faust, 2007). Before, surveys were the undisputed primary source of information on culture. The *European Values Study*

(EVS) consists of regular survey studies that trace the changes in cultural convictions of a large panel of participants in "extended" Europe. It has been conducted in four waves, every nine years: 1981, 1990, 1999, and 2008. The last study included 47 countries. The study comprises many covariates for the assessment of socio-ethnical background. The fieldwork for the most recent wave of the EVS started in September 2017. As soon as results are published, they will provide valuable insights into how the many "crises" of the past ten years affected the Europeans' values regarding religion, solidarity, and many more. Survey data and structural equation modeling aim at being as objective as possible, in a perpetual quest to find comparable yardsticks instead of subjective snapshots in space. The EVS survey implies approximately one hour of interview per person. The translation of the surveys was decentralized and harmonized between countries with the same language. Items rely on well-established scientific measurement scales. Sample sizes are counted per country, requiring 1000 participants for smaller and 1200 for larger countries. Samples are randomized to cover the social stratum of society. Information on the hierarchical structure of sample design is provided, as well as population distributions. Staff conducting interviews is thoroughly trained, and quality checks on respondents, refusals, and non-contacts are documented[1].

As Rokeach (1968) showed in the 1960's, change differs in speed for *values*, *attitudes* and *beliefs*. Values are a very long-term concept of culture, and its measurement and comparison therefore must be based on methods that show consistent results for people with similar long-lasting values across time and space. Surveys contain a meticulously chosen set of questions that have proven useful for consistent measurement of long-term stable and comparable elements of culture. Long-term oriented confirmatory approaches, based on survey data, are usually looking for objective measurement scales and therefore describe very stable, basic patterns of culture. Facebook data and search engine data are freely available and allow to compare the unobservable mindsets of people that register their nationality in the data. Who is in contact with whom can be tracked as well as short-term changes in interests and preferences. Recently, Location Based Social Networks have even made it possible

[1] `http://www.europeanvaluesstudy.eu/page/methodology-1.html`

to track the current position of people, and observe people's behavior in different contexts, for example when they are on holiday. Twitter data provides exact time stamps and geo-coordinates based on GPS, as well as information on direct interaction among people (e.g. via mutual following and re-tweets).

The translation of values into statistical terms is different from the translation of short-term opinions and attitudes. Wei et al. (2016) use Twitter data to assess the *cultural pulse* of a region, profiting from the geotagging of Twitter messages. Location Based Social Network data is highly dynamic, and thus can be used for identifying what the population is concerned with and what kind of help is needed in a crisis. The possibility of analyzing culture with such highly dynamic data raises new questions about the local embedding of culture. In virtual space, people cluster often because they are similar – not the other way around. This way, space becomes *endogenous* to its modeling, questioning the causal relationship between the two. More recent observations on human interaction in a digitized world lead to the conclusion that both the global network and the small-scale regional environment gain in importance, while national cultural imprints are becoming less important cultural exchange processes (Chua et al., 2011).

To measure long-term concepts of culture, different methods are needed, that are based on highly reliable models. For illustration, we chose two topics from the field of cultural values with high importance in the contemporaneous European context. The important increase of refugees seeking for shelter leads to a strengthening of nationalist opinions, resulting in surges in the voting outcomes of nationalist parties in France, Germany, and the Netherlands. With the recent immigration of many refugees with Muslim background, religious institutions have re-entered the focus of public discussions. An international comparison, which land's genuine population deals with religious institutions in what way can be helpful for integration of immigrants. Both the European integration as well as the integration of refugees require a high degree of solidarity. We therefore chose *solidarity* and *institutional religiousness* as the general frameworks we want to work in. Our estimations are based on the 2008 wave of the survey.

As it makes relatively complex and abstract concepts such as *religiosity* quantifiable,

literature often resorts to so-called *structural equation models* or *confirmatory factor analysis* (CFA). General scales of culture or measurement models need to be defined in a way that they are *invariant*, i.e. such that they cannot be misunderstood or measure different things across cultures. A model is said to exhibit *invariance* if it expresses the same attribute across cultures – an attribute that many emic researchers consider as non-existent. In statistical terms, invariance is present when all the parameters that are necessary to measure the construct are the same across cultures, or rather when the hypothesis that they are the same cannot be rejected. The operationalization of a single value is based on a detailed assessment of the items that this concept implies. Consider the aim of measuring "solidarity" among different cultures. Solidarity is a complex construct in itself, and surveys try to cover the visible indicators of solidarity instead of asking directly. It is possible to analyze culture in a more simple way, for example "80% of all Germans think, that..." (e.g. Halman et al., 2011; Pettersson and Esmer, 2008), but such univariate analyses are prone to misunderstandings. When using only one question at a time, there is no way to check whether all the participants of a study have understood the question in the same way, or whether one culture has a tendency to say "yes" no matter what their real conviction is. More complex constructs have the advantage that they can take account of such misunderstandings. Structural equation models take account of complex unobserved constructs via concrete survey items.

Schwartz and Huismans (1995) coin a measurement model that deals with different aspects of religiosity and the relationship of individuals towards church as an institution. This measurement model has been used, among others, by Inglehart and Baker (2000) and Halman and Riis (2003). *Institutional church religiosity* is a concept that describes the relationship between people and church as an institution rather than actual belief in god. *Church adequacy* is a sub-construct of *institutional church religiosity*, and bases on the following questions or *items*:

> Generally speaking, do you think that your church is giving in your country adequate answers to:
>
> 1. ... the moral problems and needs of the individual;
> 2. ... the problems of family life;

3. ... people's spiritual needs;
4. ... the social problems facing your country today?

(Halman and Riis, 2003, p.8)

Confirmatory latent factor models are used to describe causal relationships between unobservable, latent values or attributes (such as *church adequacy*) and observable indicators (such as answers in a questionnaire). More precisely, survey items are understood as driven by the underlying concepts of interest, formally

$$\mathbf{X} = \mathbf{M} + \mathbf{F}\mathbf{\Lambda} + \mathbf{E}. \tag{4.1}$$

Here, $\mathbf{X}$ denotes a $N \times P$ matrix of responses of person $n \in 1, \ldots, N$ to question $p \in 1, \ldots, P$, whereas the $N \times P$ matrix $\mathbf{M}$ contains the respective mean responses. Furthermore, $\mathbf{F}$ describes a $N \times K$ matrix of unobservable factors or constructs representing a cultural concept, and $\mathbf{\Lambda}$ stands for a $K \times P$ matrix of factor loadings that express the relationship between question $p \in 1, \ldots, P$ and factor $k \in 1, \ldots, K$. Error terms are included in the $N \times P$ matrix $\mathbf{E}$.

Vice versa, the observed items provide information on the latent factors. Inverting equation (4.1), unobservable constructs (such as *church adequacy*) can be expressed as linear combinations of the survey participants' answers to the related questions. Consequently, general concepts are measurable through a set of judiciously chosen queries. Moreover, every participant can be assigned a numeric value on a new, artificial scale, that assesses the individual perception of the cultural concept. Note that the confirmatory factor model requires clear hypotheses concerning which factors are represented by which items.

Subject-based literature provides numerous examples on how to embed latent factors in comprehensive causal relationships. Norris and Inglehart (2011, p.15) investigate the interplay of *religious beliefs*, *religious values*, *religious participation* and *religious political activism*, each measured by several items, in their influence on economic and social development indicators. Abela (2004) investigates the relationship of religion and spirituality with socio-economic solidarity in three different spatial scales. Basing on EVS data from the 1999 survey wave, the author combines eleven measurable

items to three indicators *local*, *social* and *global* solidarity. The eleven items follow the pattern "Are you concerned with...". Local solidarity concerns members of the immediate family, people in the neighborhood and in the community. Social solidarity describes solidarity with elderly, sick, disabled and unemployed people, whereas global solidarity treats the topic of solidarity with immigrants, humankind and Europeans (Abela, 2004).

Abela (2004) distinguishes in an empirical study using the EVS between Northern, Southern, Eastern and Western Europe. Basing on a comparison of the average score of people from these regions on the three new, artificial scales, the author concludes that people from the Mediterranean show a higher degree of social solidarity than the rest of Europe. We aggregate the data to NUTS0[2], NUTS1 and NUTS2 to illustrate the effect of regional clustering on insights. We analyze a sample of Austrian, Belgian, Czech, French, German, and Italian inhabitants that have the respective nationality and were born within the country. The results of an estimation of the level of *social solidarity*, based on local means on different aggregation scales, are illustrated in figure 4.2. Figure 4.3 shows the distribution of mean values of *church adequacy*.

Figures 4.3 and 4.2 suggest that for the 2008/2010 wave of the survey, Austria and the Czech Republic show a lower level of Social Solidarity than Belgium, Germany, and France. This is in contrast to the findings of Abela (2004), who found differences in the data from 1999/2000 in favor of a higher level of Social Solidarity in the Mediterranean region. Figure 4.3 shows, in harmony with general institution concerning religiosity, a high level of *Church Adequacy* in southern Germany, and Italy. There is a contrast between Bavaria and Upper Austria, which is also generally considered a country with high emphasis on institutional (catholic) church.

However, the figures also show that when the aggregation is on NUTS2 level only, the variation within Bavaria is just as strong as across the whole study area - showing a

[2]NUTS is French for *nomenclature des unités territoriales statistiques*, and designates a hierarchical systematization of territorial units within the European Union for standardization of statistics. NUTS0 usually designates nations, and the smaller units seek to homogeneously include the same number of individuals. A NUTS1 territory is contained in one NUTS0 territory, whereas one NUTS0 can but doesn't need to contain several NUTS1 territories. The same relationship is true accordingly for NUTS1 and NUTS2, as well as for NUTS2 and NUTS3.

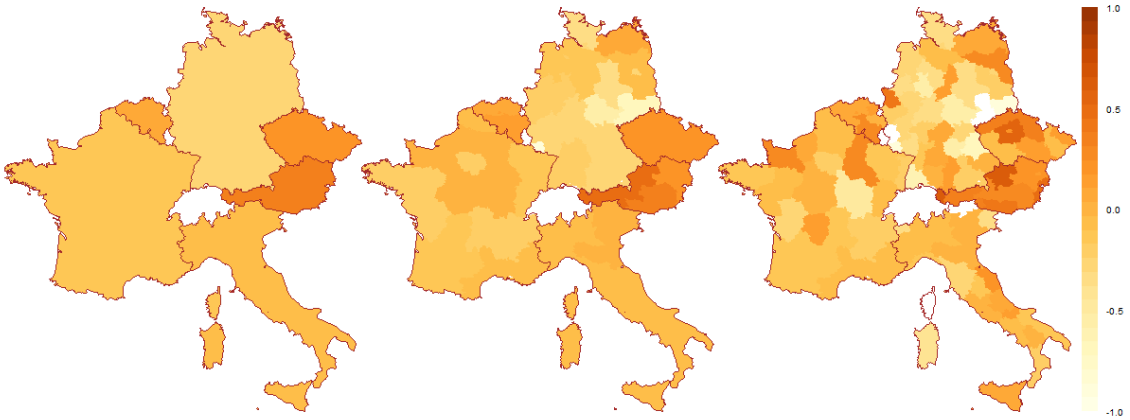

Figure 4.2: *Social solidarity* based on the 2008/2010 wave of the European Values Study, aggregated to NUTS0, NUTS1 and NUTS2

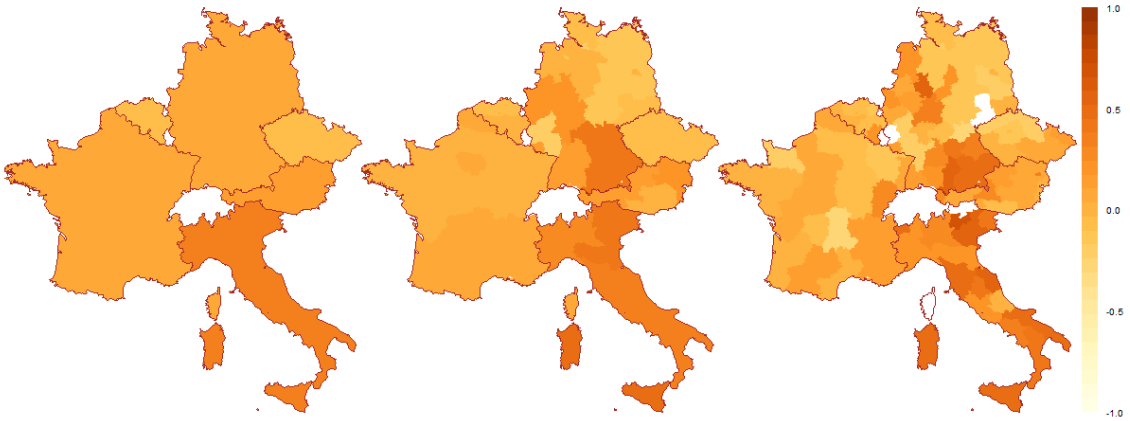

Figure 4.3: *Church adequacy* based on the 2008/2010 wave of the European Values Study, aggregated to NUTS0, NUTS1 and NUTS2.

high variance of local means within a single federal state. France, however, is very homogeneous no matter which aggregation level the means are calculated for. This might speak in favor of a long historic persistence of tradition and culture, resulting from the lack of a long common history for the single districts within Bavaria, and a long history of centralism in France. Whatever the interpretation of the values - it relies on the soundness of the measurement construct. It shows, however, that any conclusions on a national level for Germany would be dubious.

4.3.2 RQ2: Group-Based Validity Checks

In the sociological perspective on culture, the assignment of participants to groups is a central prerequisite to comparison of societies. Only if the constructs can be measured reliably across groups, and when mean values can be trusted to measure a "mean" attitude or belief, comparison is actually possible. For example, if one group has consistently higher values on the new, artificial scale of "church adequacy", members of that group can be assumed to be consistently more convinced that the church provides adequate answers to the topics mentioned in the four items. The more often a measurement scale is used, the more reliable it becomes, as usually, its adequacy is re-evaluated for every new sample.

Adopting concepts from literature without questioning their appropriateness in the research area is rightfully claimed invalid by skeptics of intercultural comparison. Poortinga (1989) early argued that scales must be validated across the populations included in a study. One of the advantages of using more than one question to measure a single concept is that it is possible to validate whether the interplay of the single questions is homogeneous across groups. The appropriate methods are implemented in any advanced statistical software. Jedidi et al. (1997) proposes a series of tests to determine the degree of homogeneity or *invariance* of the measurement models across groups. Therefore, all participants are grouped, and group-specific values for the parameters are estimated:

$$\mathbf{X}|g = \mathbf{M}^g + \mathbf{F}^g\mathbf{\Lambda}^g + \mathbf{E}^g \tag{4.2}$$

The more similar the estimated loadings, group-specific variances, and the group-specific intercepts are, the more valid is the measurement model. The determination of the degree of invariance is based on a χ^2 hypothesis testing. If the tests come to the conclusion that every group involved in the study has the same interplay of factors, factor loadings, intercept, and error structure, it can be assumed that two members from two different groups have the same perception of items and the same measurement models apply (Jedidi et al., 1997). A recent overview on testing for measurement invariance in a cross-cultural setting is provided by Milfont and Fischer (2015).

There are different forms or *levels* of invariance. A CFA with implicit group structures can measure whether the interplay between two variables that belong to the same construct is equal across two cultures (Jöreskog, 1971). This model is referred to as the *multiple group confirmatory factor analysis* (MGCFA). The *degree* of invariance of a measurement scale can vary - some concepts might have approximately the same meaning across cultures, but differ in some details. One of the core advantages of assessing a cultural concept not via a single question, but with several questions as given in the examples described above, is that the interplay of the single items can be checked for consistency across groups.

The step-wise validation of measurement invariance proposed by Jedidi et al. (1997) detects problems in the measurement construct. These "problems" can be of different nature. Concept-related problems can appear if, for example, the understanding of a concept is different among groups. Depending on the part of the model that is affected by a lack of invariance, the tests can come to the conclusion that a model's parameters show *construct*, *structural* (or *functional*), *metric* (or *measurement unit*), and *scalar* (or *full score*) equivalence (Van de Vijver, 2011, pp.6-9). For example, going to church is in some countries closely related to being religious, and in others not. A graphical illustration for the situation of scalar, metric and structural invariance can be found in figure 4.4, which illustrates the situation that group membership affects the parameters in the model. In the middle panel for example, group membership affects the items. This is the case for example for an acquiescence bias, where people from a single culture have a tendency to answer questions with

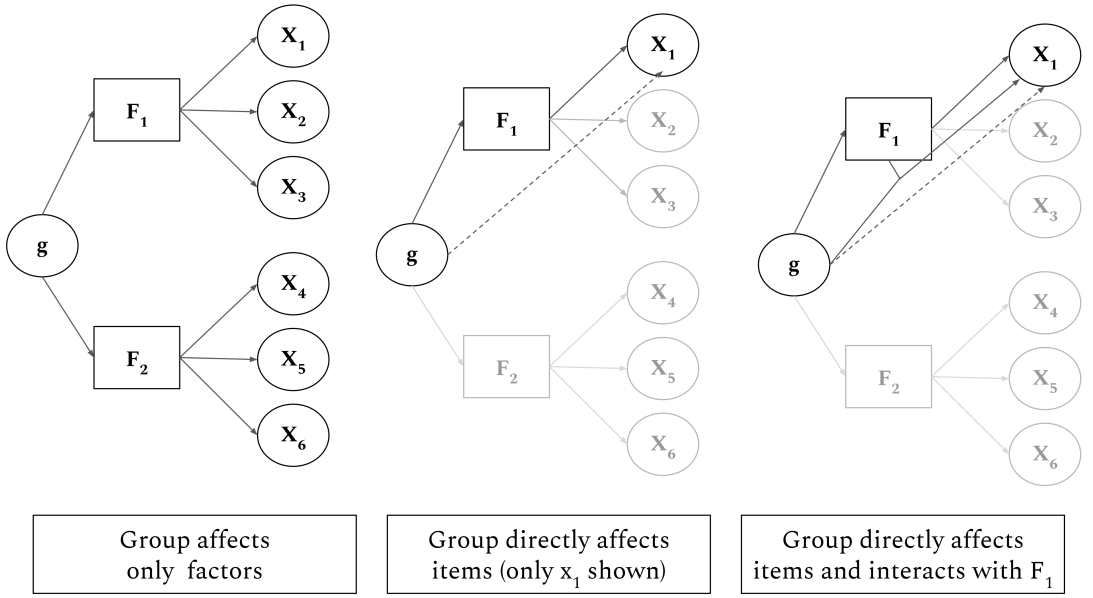

Figure 4.4: Visualization of the group-based structural equation model. F_1 and F_2 affect three observable items, respectively. The degree of invariance decreases from left to right, and correspond to *scalar*, *metric* and *structural* invariance.

"yes" rather than "no" because they don't like to say "no", irrespective of their factor values. The third graphic in figure 4.4 means that the factorial structure is different, i.e. the item loadings differ from group to group, additional to the effect on the item itself.

Following a slightly different categorization system, Kankaraš (2010) subsume this kind of problem under the label "construct bias". A "method bias" reveals itself not within the content of a concept, but in its measurement. An "acquiescence bias" as a special case of the "method bias" describes the situation when members of a group tend to prefer answering "yes" to a question rather than "no". More examples can be found in Kankaraš (2010) and Van de Vijver (2011). The information on which part of the model diverges most between two groups can be analyzed via so-called *modification indexes*. These provide information on which group behaves different in what aspects, and where the largest potential for wrong conclusions lies (see Wu et al., 2007). Such findings can be interesting also from an emic perspective as modification indexes can provide guidance to detect differing behavior.

Ignoring the potential lack of invariance between groups can lead to substantially

wrong conclusions. According to Dülmer (2011), there is a lack of invariance in the measurement models of Inglehart and Baker (2000) and Norris and Inglehart (2011), which leads to invalid conclusions on the relationship between economic development and social values and norms. Scheepers et al. (2002), not validating measurement invariance, find a positive relationship between religiosity and ethnic prejudices. Cambré et al. (2002) come to the opposite conclusion after correcting for a measurement bias, due to the positive formulation of some items in the survey the study bases on. This means that in some languages, items were translated such that they allowed people to answer with "yes" to express the same opinion, a case of the mentioned acquiescence bias (see also Moors, 2004; Savalei and Falk, 2014). Beuckelaer and Swinnen (2012) conduct a simulation study, that illustrates the consequences of a deviance in single parameters between groups. The problem of lack of invariance is also relevant for the example we use: Kankaraš and Moors (2008) replicate the study conducted by Abela (2004) and verify invariance of the model between countries. The authors come to the conclusion that the concept *social solidarity* is not invariant, and that the lack of invariance influences the results of Abela (2004).

Fit-indicators such as the *comparative fit index* (CFI), the *Tucker Lewis index* (TLI) and the *root mean squared error of approximation* (RMSEA) correct for sample size and should therefore be used instead of χ^2 hypothesis tests (for a discussion see also Meuleman and Billiet, 2012; Hirschfeld and von Brachel, 2014). Cut-off limits for very good, good and mediocre fit for the RMSEA are, according to MacCallum et al. (1996) 0.01, 0.05 and 0.1, respectively. Moreover, CFI $\geq$ 0.90 or 0.95 and RMSEA $\leq$ 0.06 or 0.08 were proposed (Wu et al., 2007). As Wu et al. (2007) point out, the decision rules for such criteria are still sensitive with regard to model complexity. As our models are fairly simple, involving no relationships between constructs and basing on a limited number of items, the rules we apply should be rather strict. Decision criteria have to be adapted to sample size and model complexity – for example when the constructs interact with one another, or are part of a causal model chain. For implementation, we use the package `lavaan` (Rosseel, 2012), which provides an interface for group-based structural equation modeling.

Table 4.1: Fit indexes and χ^2 tests for measurement invariance for mixed data, for the four-factor model comprising three kinds of solidarity measures and *church adequacy*.

Inv.	CFI	TLI	RM-SEA	Df	χ^2	Δ Df	Pr($>\chi^2$)
config.	0.985	0.107	0.107	426	6 257		
weak	0.975	0.973	0.127	496	10 091	8.945	$< 10^{-15}$
strong	0.971	0.975	0.123	626	12 009	13.903	$< 10^{-15}$

When confronted with binary or ordinal data, as in our case, the function switches from maximum likelihood estimation to *diagonally weighted least squares* (DWLS) (Muthén, 1993). This approach uses the weighted least squares estimator and a polychoric correlation matrix as input.

Using the 2008 data, we tested for invariance on NUTS0, i.e. the national level. In general, the models provide a reasonable fit, as the χ^2-based tests show. Table 4.1 shows our results, indicating that both fit-based and χ^2-based tests reject strong invariance. Testing for invariance on a national, i.e. NUTS0 level, reveals that the largest difference in the model parameters manifest for the Czech Republic. The direct comparison of the estimated parameters shows that answer behavior differs from the other countries for the questions of adequate answers to *family life problems* and *spiritual needs*. Modification indexes also indicate that there is a lack of invariance with regard to the question "Are you concerned about your immediate family?". The results indicate that group membership has a strong relationship with this covariate, which is not entirely mediated through differences in *local solidarity*. As shown above, the group-based confirmatory factor analysis seeks for a maximum of objectivity by prescribing concrete validation steps. Participants are embedded in groups *before* the models are estimated, and invariance can only be tested across those groups. The spatial scale can be adapted to the solicited degree of detail – still, only the *a priori* assignment to groups allows to separately estimate parameters and check for invariance.

Many examples of empirical research accept nationality without further discussion as a proxy for cultural collectivity. Moreover, most empirical research chooses the group

assignment very pragmatically, often with direct reference to nationality (Beuckelaer and Swinnen, 2012). Moors (2004) admit the pragmatic decision path in their investigation of invariance:

> Of course any covariate can be used as a group variable, but since construct equivalence is analyzed within the context of cross-cultural comparison, the covariates should at least make partial reference to the identification of cultural groups (Moors, 2004, p. 307).

The author chooses ethnicity as a grouping variable and, with reference to possible mistakes in survey translation, the language in which the survey was conducted with the participant. Measurement models that allow for heterogeneous parameter values among groups seek to offer a tool to check the comparability of two mean values across regions. When hypothesis tests on the invariance of models do not reject the homogeneity of parameters, it can be considered safe to use a measurement model for cross-cultural comparison. If it is rejected, then there is a strategy to check the single parameters and identify sources of misfit (compare Steiger, 1990). The structural equation approach to the modeling of culture is *etic* in nature (i.e. it applies a meta-perspective on culture), and constructs that are used within such a procedure must be well-established to ensure they are objective. It is necessary to group individuals into their cultural affiliation before applying the method.

4.3.3 RQ3: The Consequences of Repeating National Blueprints

The ever repeated usage of the same measurement models, basing on years of validation, is disputed by emic researchers, who doubt that the establishment of *objective* measurement scales is feasible and useful. Moverover, the large discretionary scope for decisions on group affiliation has incited some defenders of quantitative research to dispute the importance of invariance tests. Beuckelaer and Swinnen (2012) argue that most measurement models can almost never be classified as invariant across groups, and therefore invariance tests should not be applied in an overly rigid way. However, as the discussion of the consequences of group-based comparison without invariance validation has shown, neglecting the tests is very risky. The MGCFA

impedes the erroneous application of non-invariant models. Via modification indexes, which show which items cause the highest inconsistencies between groups, the MGCFA also provides important clues of where the problem might lie.

There are many approaches in the literature that try to infer the homogeneity of groups without fixing groups prior to model building. Jedidi et al. (1997) investigate consequences of separating individuals into groups only *after* parameter estimation via cluster methods. First, the model parameters are inferred without grouping the sample, and every participant is assigned a factor score. Second, people are clustered into variable regions, with relatively homogeneous factor values. Third, the clustered regions are tested for measurement invariance in the same procedure as described above. Jedidi et al. (1997) show via simulation that this procedure leads to biased results. The problem consists in the fact that the model that the grouping bases upon is already misspecified, and parameter estimates are already biased in the first step. When the assignment of people to regions is already based on a misspecified model, then subsequent tests can be meaningless, especially as group assignment is endogenous to the same sample of values that the invariance tests base upon.

Furthermore, statistical methodology for parameter inference is an issue. As Hjort and Claeskens (2003) explain, the usage of frequentist models results from the high computational costs that Bayesian analysis brought along and that used to be prohibitive. The frequentist school of thought, which is denoted "traditional" by Hjort and Claeskens (2003), ignores the fact that model specification has been proceeded by many rounds of model specification and validation. Surveys are conducted in a way that reproduces invariant models across cultural groups and nations, and the ever-repeated focus on nationality has led to a selection of models that reproduce national differences.

Probabilistic Bayesian approaches allow to include uncertainty measures and subjective beliefs in a model, that can then compete against other beliefs. In empirical research, model selection strategies are commonplace. However, when the most fitting, or best performing model has been identified, model specification and estimation proceeds as if the model had been known in advance. Often, this leads to too

optimistic confidence intervals and the underreporting of uncertainty. In *frequentist* terms, parameters have no distributions, but only a true value, which is estimated via a supposedly random sample. Therefore, the approach does not distinguish between prior and posterior information. Point estimates are used and inference is conducted under the assumption that the model used for estimation is the *true* model. This means that the complexity of reality is accounted for by inference that is only valid within the given data set under a very ambitious assumption. Bayesian approaches quantify their prior beliefs in the form of prior distributions of parameters. In Bayesian models, prior beliefs and the likelihood of observing the given data merge in a posterior distribution. The underlying assumptions are ambitious in a different way, as prior distributions have to be formalized such that they can be assumed to be true or at least not detrimental if there is only few prior knowledge. To reconcile both perspectives and to assess the consequences of model selection prior to computation, Hjort and Claeskens (2003) base upon a concept that both schools of thought are familiar with: likelihood. The authors propose a *competing model average* approach, where several competing models enter together.

4.3.4 RQ4: How to Model Cultural Exchange Processes

The MGCFA assumes that within a pre-defined group, variance-covariance structures between factors and items are homogeneous, i.e. that the same relationships exist between factors and items. The model assigns each observation to exactly one group, and for the estimation of group-specific parameters, only the samples assigned to this group are available. If the number of survey participants is small, this will lead to issues with sample size, and a lack of power of statistical tests (i.e. the probability that a false hypothesis is rejected is small). Invariance tests can only be conducted as long as the group sizes are large enough. It is possible to conduct the invariance tests on larger regional levels, and then aggregate the values for smaller regions – leading to the more fine-grained coloring of the countries. Nevertheless, each group is treated as independent from one another.

So what about approaches that specify a spatial exchange of culture? Can this cultural exchange be taken into account? There are approaches that specify dependen-

cies and interaction on different levels of a model, allowing to consider small regions and their national embedding at the same time. Variability in survey responses can be represented as factors that exert influence on a local, regional, national, or even supra-national level. Methodological bases for this kind of analysis in space have been developed by Wang and Wall (2001, 2003). Hogan and Tchernis (2004) describe a Bayesian hierarchical model for factor analysis of data with a spatial dependency structure. The authors analyze spatially correlated multivariate data reflecting the distribution of material deprivation across the U.S., more specifically for a cross section of census tracts in Rhode Island. People's unobserved material deprivation manifests in several observed categories of non-monetary poverty. It is assumed that there is one common latent variable that describes a local unobservable factor. This means that the factor scores reflect a snapshot of the actual state of poverty disparities, with regard to different dimensions of deprivation.

The situation that the authors depart from is fundamentally different from the aims we are following above. The data, like in our case, consists of four indicators of a latent construct – i.e. material deprivation. However, the aim is not to compare regions and find statistically significant differences - as for every region, there will be only one factor value, and neither would a statistical test make sense on the basis of four highly correlated values. The authors rather seek to find a good linear combination of deprivation indicators, that describes the situation well. For this purpose, the authors suggest to specify a spatial dependency between the indicators. There are several *levels* in the model, which is the reason for it being called *hierarchical*. In the *first level* of the model, there is a single factor of deprivation, and every region has its own factor value, i.e. a certain level of deprivation. On the *second level* of the model, the area-specific factors have a joint distribution, respecting prior hypotheses on the pair-wise relationship (*neighborhood*) of particular regions. Therefore, a central step in hierarchical modeling is the definition of neighbors - which regions are neighbors to one another? Neighborhood can be defined *ad hoc*, for example as all regions that have a common border with the region of interest. However, neighborhood can also be defined basing on a prior hypothesis, such as the hypothesis that former Soviet states can be neighbors to one another, but the late consequences

of the iron curtain nowadays still culturally separates the East from the West. A model can be specified such that no neighborhood beyond a national border is possible. The core advantage of the described procedure is that it does not necessarily require an *ad hoc* specification of neighbors, rather, the model can be used to decide which setting best describes the observed data structure. By measuring the fit of a model, it is possible to decide on a likely neighborhood structure.

To operationalize the model, a hypothesis has to be found on what spatial correlation depends on. Wang and Wall (2003) for example assume the spatial covariance matrix to depend on a measure of neighborhood of areas, more precisely on whether two areas share a common boundary, and specify the assorted *conditional autoregressive model*. This again involves a prior hypothesis on neighborhood, and therefore doesn't resolve the need for prior assumptions on cultural exchange processes and collectivity. However, it resolves the problem of sample size by including information from neighboring areas into the estimation of factor structures. With regard to testing the invariance of measurement models, to the best of our knowledge there is no application that uses fixed spatial correlation structures in model parameter invariance testing.

The core difference to the models discussed above is that under consideration of a spatial dependency structure, the estimation of the latent factors that represent material deprivation or church adequacy does not rely on the improbable assumption that the two groups are independent. At the same time, it also changes the logic of quantitative cultural research: If the two countries are not assumed independent, it makes less sense to compare them via statistical tests like testing the equality of means. Moreover, in the case discussed by Hogan and Tchernis (2004), there are very few observations, and the specification of a spatial dependency structure stabilizes the variance of the results by using some neighboring information in the estimation of the factor values. In our case, we have many observations. For the case of 1000 participants in a survey, a neighborhood matrix would contain 10^6 elements – which makes the estimation of the latent factorial structure pretty tedious.

4.4 Discussion

4.4.1 Sociological Group-based Mapping of Culture

We proposed four research questions with regard to the statistical modeling of culture and applied various models to an example data set from the European Values Study. All four research questions depart from the presumption that culture is homogeneous within groups, and hypothesize that researchers want to find out about the spatial distribution of cultural beliefs and values. In the following, we discuss the main insights from the models presented above.

Cultural concepts like values or attitudes cannot be observed directly. Therefore, measurement models that assume that these underlying concepts manifest in multidimensional survey answers are needed to provide reliable evidence on culture. To answer RQ1 (How can cultural collectivity be quantitatively assessed and compared across different regions?), we have introduced structural equation models, that solve the problem by applying measurement models which have proven to be reliable in previous exploratory and confirmatory studies. Fit measures assure that the model sufficiently well describes the correlation structure of the data. These measurement models can be used to make unobservable concepts observable.

With regard to intercultural comparison, the way the results are used is misguiding. Often, the discussion of a cross-cultural comparison study is limited to the calculation of a regional mean and the coloring of maps representing political units. This creates an impression of homogeneity where it is unlikely to exist. It is, however, the only assumption that allows the verification of models. Despite the many difficulties arising in cross-cultural comparison, a differentiated and context-specific cross-cultural analysis can be a powerful tool to gain insights into the thinking schemes, convictions, and temporary attitudes of people from different regions. The general insight that can be gained from the approaches above is that in order to conduct a cross-group comparison, groups must be assigned *a priori*, and therefore in order to use the model, researchers must accept that *group structures precede culture*.

In order to ensure that a comparison is valid, it has to be investigated whether

the measurement models are describing the same underlying unobserved cultural concepts across groups. Based on the a priori assignment of groups, it is possible to verify the quality of models. In MGCFA, the analysis of the objectivity of a measurement model lies at the center of interest. It thus provides an answer to our RQ2 (Which statistical methods and models can ensure comparability of different measurements?). In matters of content, but also with regard to unbiased parameter inference, it is important to accept group assignments as given. Problems concerning the automatic rejection of measurement invariance for large samples have created skepticism among empirical researchers about the relevance of tests. However, simply ignoring invariance tests has been shown to bear important risks of erroneous conclusions.

This problem regarding test reliability, as well as the possible consequences of the ever-repeated application of the same models, which have "proven" to be reliable, have given rise to RQ3 (How can the consequences of an endless repeating of national blueprints be translated into statistical terms?). The discussion shows that when always relying on the same blueprints for grouping people into cultural groups can lead to problems with model validation, as researcher may single out specific models that meet their expectation with regard to outcomes over time.

Our discussion of RQ4 (How can the model be specified with more flexibility with regard to grouping?) emphasizes the inclusion of parts of information from other areas, thus allowing to be more fine-grained with regard to areas defining cultures. The methods discussed in RQ1 to RQ3 all accept the assignment of all individuals to exclusive groups. Especially in federal systems, it is however likely that individuals belong to more than one group, or that their group is part of a larger group (e.g., Franconia is part of Bavaria, which is part of Germany). Giving up the notion of groups would impede the discussion of cultural comparison – when a comparison is of interest, criteria that determine whether someone is part of a culture are needed. If more flexibility is warranted without giving up the notion of cultural collectivity, models that specify hierarchical levels of spatial relationships can be a good solution. Hierarchical models assume a specific correlation structure for the latent factors associated with the lower-level regions. That means, there is spatial correlation that

relates neighboring areas. One consequence is that the group size problem which occurs in the models described for RQ2 is alleviated by using information from neighboring areas for estimating the parameters of one subgroup. In the meantime, this approach has also been extended to spatio-temporal settings, where changes in the observed data over time are explained by an interplay of spatial and temporal factors (compare Bruno et al., 2013).

Latent factors can be integrated into a larger model, allowing the analysis of the interplay of different factors, and the inclusion of exogenous predictors, that seek to explain the observed structure (compare Liu et al., 2005). This means that it is possible to thoroughly account for spatial processes, and include covariates like nationality in the model, in order to conduct thorough inference for whether nationality really makes a difference for unobserved cultural values (e.g. solidarity). Then, spillover effects across a border do not influence the results. There is, however, a core difference to the model discussed above: While it involves a specification of spatial proximity and neighborhood, it does not involve parameter testing, and there is no explicit structural equation modeling. This makes the testing of parameter equality redundant.

The relevance of clear-cut borders for the assessment of culture across Europe has been shown both from a content and from a methodological perspective. A spatial understanding of long-term cultural concepts like *values* accepts that *geography precedes culture*. For more short-term concepts, data-driven approaches are more likely to provide interesting insights into cultural dynamics. Values, being more fundamental, require *objective* measurement scales and clear-cut a priori assignment to groups. The denomination of fixed borders that delineate homogeneous groups from one another is necessary for checking the validity of a measurement model. There is, however, no statistical reason to stick to national borders for this purpose. Rather, any quantitative analysis should be preceded by an analysis of probable group structures.

4.4.2 Learnings for Alternative Paradigms in Cultural Research

The curated dataset that the EVS provides is limited in its ability to represent changes in the data due to the high costs of data production. More versatile data sources have been used for cultural comparison in the last years - such as social media datasets. Even though the models and approaches used in that context make extensive use of statistics and probabilities, involving mathematical equations and explicit definition of relationships between individuals, as well as machine learning for optimizing prediction accuracy and parameter inference.

The hypothesis-based, statistical approach that we discussed in the previous sections has been shown to depend on blueprints and clearly formulated, statistically verifiable hypotheses. While recently developed approaches also make use of statistical models, the nature of their uncurated and unstructured data may make the clear distinction of "big data" from small data necessary. Manovich (2018) even suggests that the difference may be as large as the difference between statistical models and deterministic laws on human behavior and simulation-based research. Because of the unlimited number of features that can be measured for every single human, the level of detail of analysis is unbounded, and offers new perspectives on what culture is and how a society can be described, digging much deeper into cultural details that the model-based approaches described above.

In many applications, approaches based on new data sources can enhance our understanding of culture. For example, social network data can be used to measure cultural diffusion processes. In a combination of big data, simulation methods, and statistical tests, Axelrod (1997) showed how cultural segregation can persist in a globalized society, with a model-based simulation on attracting and detracting forces among culturally more and less similar people, putting a hypothesis of homophily – that is, the tendency of humans to engage with similar rather than different humans – into the model. This has important implications for the co-existence of cultural majorities and minorities and political decisions that influence this co-existence (see also Flache and Macy, 2011). Computational social science, when taking advantage of the large datasets that are available today, must keep the advantages and implications of single approaches in mind. Multiple research paradigms can be com-

bined to gain insights into society and to inform decision makers about the potential consequences of decisions (Conte et al., 2012).

4.5 Conclusions

We have shown that inter-cultural comparison is related to the definition of clear-cut cultural boundaries, both from socio-theoretical and a statistical perspective. Recurring to the rhetoric of clear-cut cultural boundaries is more than ever popular for the simplified mediation of cultural differences, but at the same time ever further from the reality of globalizing cultural exchange processes. We are in between the two extremes of the range of opinions. We showed that the separability of collectives can be done in a context-specific way, avoiding the dominance of national blueprints for cultural analysis. We modeled different forms of solidarity and church adequacy and showed that national patterns were to measurable, but that within-country variance dominated the across-country variance. This local variance holds especially in countries with relatively short history as a nation. On the one hand, these patterns show that there is no empirical basis for the belief that cultural clusters are inevitably congruent with political borders or with one another across different contexts.

Questioning the meaningfulness of political borders does not necessarily lead to the denial of regional grouping as a tool for mediation of cross-cultural differences, as long as the grouping is understood as a flexible categorization, tailored to the specific interest regarding a single cultural concept. The European context, rich in cross-border history, offers many reference points for the reformulation of borders. Consequently, every single research project requires an elaborate empirical validation. However, only this costly procedure allows us to discuss similarity as well as difference across countries.

Accepting the sociological mapping of culture to pre-defined groups, we aimed to model coherent regions of homogeneous convictions. Tests for structural equivalence showed that while there are relatively strong national differences for the solidarity concept, there are few in the religiosity example. This can provide some guidelines for future research. Taking into consideration the risk of incurring sample bias

and abandoning objectivity in data driven approaches, we advocate the usage of structural equation based, confirmatory methods as an important tool for spatial culture investigation. The adequacy of the blueprint that is currently used should always be questioned, as grouping by nationality is only one of many possibilities. When applying models to uncurated, large datasets, the drawbacks of these data sets with their undisclosed biases and many errors should be investigated and made transparent.

4.6 References

Abela, A.M., 2004. Solidarity and religion in the European Union: A comparative sociological perspective, in: Xuereb, P. (Ed.), The Value(s) of a Constitution for Europe. European Documentation and Reserach Centre, University of Malta, 71–101.

Ahern, K.R., Daminelli, D., Fracassi, C., 2015. Lost in translation? The effect of cultural values on mergers around the world. Journal of Financial Economics 117, 165–189. doi:10.1016/j.jfineco.2012.08.006.

Anderson, P., Keil, S., 2016. Minority nationalism and the European Union. The cases of Scotland and Catalonia. L'Europe en Formation 379, 40–57. doi:10.3917/eufor.379.0040.

Ang, S., Van Dyne, L., 2015. Preface and acknowledgements, in: Ang, S., Van Dyne, L. (Eds.), Handbook of Cultural Intelligence. Taylor & Francis, XV–XVIII.

Antonczyk, R.C., Salzmann, A.J., 2014. Overconfidence and optimism: The effect of national culture on capital structure. Research in International Business and Finance 31, 132–151. doi:10.1016/j.ribaf.2013.06.005.

Axelrod, R., 1997. The dissemination of culture: A model with local convergence and global polarization. Journal of Conflict Resolution 41, 203–226. doi:10.1177/0022002797041002001.

Barocas, S., Hood, S., Ziewitz, M., 2013. Governing Algorithms: A Provocation Piece. Technical Report. Available at SSRN: https://ssrn.com/abstract=2245322. doi:10.2139/ssrn.2245322.

Benedikt Gräler, E.P., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. The R Journal 8, 204–218.

Bernaards, C.A., Jennrich, R.I., 2005. Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. Educational and Psychological Measurement 65, 676–696.

Berry, D., 2011. The computational turn: thinking about the digital humanities. Culture Machine 12.

Beuckelaer, A.D., Swinnen, G., 2012. Biased latent variable mean comparison due to measurement noninvariance: A simulation study, in: Davidov et al. (2011). Chapter 5. 117–148.

Bivand, R., 2017. `classInt`: Choose Univariate Class Intervals. URL: `https://CRAN.R-project.org/package=classInt`. `R` package version 0.1-24.

Bivand, R., Keitt, T., Rowlingson, B., 2017. `rgdal`: Bindings for the Geospatial Data Abstraction Library. URL: `https://CRAN.R-project.org/package=rgdal`. `R` package version 1.2-7.

Bivand, R., Pebesma, E., Gómez-Rubio, V., 2013. Applied Spatial Data Analysis with R. Use R!, Springer, New York, USA.

Bruno, F., Cocchi, D., Paci, L., 2013. A practical approach for assessing the effect of grouping in hierarchical spatio-temporal models. AStA Advances in Statistical Analysis 97, 93–108. doi:`10.1007/s10182-012-0193-6`.

Brunsdon, C., Chen, H., 2014. `GISTools`: some further GIS capabilities for `R`. URL: `https://CRAN.R-project.org/package=GISTools`. `R` package version 0.7-4.

Cambré, B., Welkenhuysen-Gybels, J., Billiet, J., 2002. Is it content or style? An evaluation of two competitive measurement models applied to a balanced set of ethnocentrism items. International Journal of Comparative Sociology 43, 1–20. doi:`10.1177/002071520204300101`.

Chua, V., Madej, J., Wellman, B., 2011. Personal communities: the world according to me, SAGE Publications, London, UK. Chapter 8, 101–115. doi:`10.4135/9781446294413.n21`.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M., Helbing, D., 2012. Manifesto of computational social science.

The European Physical Journal Special Topics 214, 325–346. doi:10.1140/epjst/e2012-01697-8.

Davidov, E., Schmidt, P., Billiet, J., 2011. Cross-Cultural Analysis: Methods and Applications. European Association of Methodology Series, Taylor & Francis.

Delanty, G., 2003. Europe and the idea of "Unity in Diversity", in: Lindahl, R. (Ed.), Whither Europe? Borders, Boundaries, Frontiers in a Changing World. CERGU, Gothenburg.

Dülmer, H., 2011. A multilevel regression analysis on work ethic as a two-level latent dependent variable, in: Davidov et al. (2011). 311–340.

Epskamp, S., 2014. semPlot: Path diagrams and visual analysis of various SEM packages' output. URL: http://CRAN.R-project.org/package=semPlot. R package version 1.0.1.

Everitt, A., 2006. Europe: United or Divided by Culture. Institute of Welsh Affairs.

Faust, K., 2007. Very local structure in social networks. Sociological Methodology 37, 209–256. doi:10.1111/j.1467-9531.2007.00179.x.

Featherstone, M., 1995. Undoing culture: Globalization, postmodernism and identity. Volume 39. Sage, London, Thousand Oaks, New Delhi.

GESIS Data Archive for the Social Sciences, 2018. EVS Longitudinal Data File 1981-2008. doi:10.4232/1.12253.

Gesteland, R.R., 2002. Cross-cultural business behavior: Marketing, negotiating, sourcing and managing across cultures. Technical Report. Copenhagen Business School.

Halman, L., Riis, O., 2003. Religion in a Secularizing Society: The Europeans' Religion at the End of the 20th Century. European Values Studies, Brill.

Halman, L., Sieben, I., van Zundert, M., 2011. Atlas of European Values. Trends and Traditions at the turn of the Century. European Values Studies, Brill.

Harrell Jr, F.E., 2017. Hmisc: Harrell Miscellaneous. URL: https://CRAN.R-project.org/package=Hmisc. R package version 4.0-3.

Helfrich, H., 2003. Methodologie kulturvergleichender psychologischer Forschung, in: Thomas, A. (Ed.), Kulturvergleichende Psychologie. 2 ed.. Hofgrefe, Göttingen, GER, 111–138.

Hermans, H.J.M., Kempen, H.J.G., 1998. Moving cultures: The perilous problems of cultural dichotomies in a globalizing society. American Psychologist 53, 1111–1120. doi:10.1037/0003-066x.53.10.1111.

Hirschfeld, G., von Brachel, R., 2014. Multiple-group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal indicators. Practical Assessment, Research and Evaluation 19, 1–12.

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879–899. doi:10.1198/016214503000000828.

Hofstede, G., 2001. Lokales Denken, globales Handeln. Interkulturelle Zusammenarbeit und globales Management. Deutscher Taschenbuchverlag, Munich, GER.

Hofstede, G., 2002. Dimensions Do Not Exist: A Reply to Brendan McSweeney. Human Relations 55, 1355–1361. doi:10.1177/00187267025511004.

Hofstede, G., 2003. Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. Sage publications, Thousand Oaks, USA.

Hogan, J., Tchernis, R., 2004. Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. Journal of the American Statistical Association 99, 314–324. doi:10.1198/016214504000000296.

Inglehart, R., Baker, W.E., 2000. Modernization, cultural change, and the persistence of traditional values. American Sociological Review 65, 19–51. doi:10.2307/2657288.

Jedidi, K., Jagpal, H.S., DeSarbo, W.S., 1997. Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity. Marketing Science 16, 39–59. doi:10.1287/mksc.16.1.39.

Jöreskog, K., 1971. Simultaneous factor analysis in several populations. Psychometrika 36, 409–426. doi:10.1007/BF02291366.

Kahle, D., Wickham, H., 2013. ggmap: Spatial visualization with ggplot2. The R Journal 5, 144–161.

Kankaraš, M., 2010. Essays on Measurement Equivalence in Cross-Cultural Survey Research. A Latent Class Approach. Ph.D. thesis. Tilburg University.

Kankaraš, M., Moors, G., 2008. Measurement Equivalence and Extreme Response Bias in the Comparison of Attitudes across Europe. IRISS Working Paper Series 2008-06. IRISS at CEPS/INSTEAD. doi:10.1027/1614-2241/a000024.

Kirby, J.B., Bollen, K.A., 2009. Using instrumental variable tests to evaluate model specification in latent variable structural equation models. Sociological Methodology 39, 327–355. doi:10.1111/j.1467-9531.2009.01217.x.

Kitchin, R., 2017. Thinking critically about and researching algorithms. Information, Communication & Society 20, 14–29. doi:10.1080/1369118X.2016.1154087.

Kitchin, R., P. Lauriault, T., 2015. Small data in the era of big data. GeoJournal 80, 463–475. doi:10.1007/s10708-014-9601-7.

Kraidy, M., 2005. Hybridity, or the cultural logic of globalization. Temple University Press. doi:10.26530/oapen_626979.

Liu, X., Wall, M.M., Hodges, J.S., 2005. Generalized spatial structural equation models. Biostatistics 6, 539–557. doi:10.1093/biostatistics/kxi026.

MacCallum, R.C., Browne, M.W., Sugawara, H.M., 1996. Power analysis and determination of sample size for covariance structure modeling. Psychological Methods 1, 130–149. doi:10.1037/1082-989x.1.2.130.

Manovich, L., 2018. The science of culture? social computing, digital humanities and cultural analytics. URL: osf.io/preprints/socarxiv/b2y79, doi:10.22148/16.004.

Martin, J.N., Nakayama, T.K., 2010. Intercultural communication in contexts. McGraw-Hill Higher Education, New York.

Masson, E., Schäfer, T., van Es, K., 2017. Humanistic Data Research: An Encounter between Academic Traditions. Amsterdam University Press. 25–37.

McSweeney, B., 2002. Hofstede's model of national cultural differences and their consequences: A triumph of faith-a failure of analysis. Human Relations 55, 89–118. doi:10.1177/0018726702551004.

Meuleman, B., Billiet, J., 2012. Measuring attitudes toward immigration in Europe: the cross-cultural validity of the ESS immigration scales. ASK Research & Methods 21, 5–29.

Milfont, T.L., Fischer, R., 2015. Testing measurement invariance across groups: Applications in cross-cultural research. International Journal of Psychological Research 3, 111–130.

Miller, H., 2010. The data avalanche is here. shouldn't we be digging? Journal of Regional Science 50, 181–201. doi:10.1111/j.1467-9787.2009.00641.x.

Moors, G., 2004. Facts and artefacts in the comparison of attitudes among ethnic minorities: A multigroup latent class structure model with adjustment for response style behavior. European Sociological Review 20, 303–320. doi:10.1093/esr/jch026.

Moran, R.T., Abramson, N.R., Moran, S.V., 2014. Managing cultural differences. Routledge, London, UK, and New York, USA.

Muthén, B.O., 1993. Goodness of fit with categorical and other nonnormal variables, in: Bollen, K., Long, J. (Eds.), Testing structural equation models. Sage Publications. Volume 154, 205–234.

Neuwirth, E., 2014. RColorBrewer: ColorBrewer Palettes. URL: https://CRAN.R-project.org/package=RColorBrewer. R package version 1.1-2.

Norris, P., Inglehart, R., 2011. Sacred and Secular: Religion and Politics Worldwide. Cambridge Studies in Social Theory, Religion and Politics, Cambridge University Press. doi:10.1017/cbo9780511894862.

Pettersson, T., Esmer, Y., 2008. Changing Values, Persisting Cultures: Case Studies in Value Change. European Values Studies, Brill.

Poortinga, Y.H., 1989. Equivalence of cross-cultural data: an overview of basic issues. International Journal of Psychology 24, 737–756. doi:10.1080/00207598908247842.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, AT. URL: http://www.R-project.org/.

Revelle, W., 2014. psych: Procedures for Psychological, Psychometric, and Personality Research. URL: https://CRAN.R-project.org/package=psych. R package version 1.4.8.

Rippl, S., Seipel, C., 2008. Methoden kulturvergleichender Sozialforschung. Springer, Wiesbaden, GER. doi:10.1007/978-3-531-19969-6.

Rokeach, M., 1968. Beliefs, Attitudes, and Values: a Theory of Organization and Change. Jossey-Bass, San Francisco, US.

Rosseel, Y., 2012. lavaan: An R package for structural equation modeling. Journal of Statistical Software 48, 1–36. doi:10.18637/jss.v048.i02.

Sarkar, D., Andrews, F., 2016. latticeExtra: Extra Graphical Utilities Based on Lattice. URL: https://CRAN.R-project.org/package=latticeExtra. R package version 0.6-28.

Savalei, V., Falk, C.F., 2014. Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. Multivariate Behavioral Research 49, 407–424. doi:10.1080/00273171.2014.931800.

Scheepers, P., Gijsberts, M., Hello, E., 2002. Religiosity and prejudice against ethnic minorities in Europe: Cross-national tests on a controversial relationship. Review of Religious Research 43, 242–265. doi:10.2307/3512331.

Schwartz, S.H., Huismans, S., 1995. Value priorities and religiosity in four western religions. Social Psychology Quarterly 58, 88–107. doi:10.2307/2787148.

semTools Contributors, 2016. semTools: Useful tools for structural equation modeling. URL: https://CRAN.R-project.org/package=semTools. R package version 0.4-14.

Snijders, T.A.B., 2001. The statistical evaluation of social network dynamics. Sociological Methodology 31, 361–395. doi:10.1111/0081-1750.00099.

Steiger, J.H., 1990. Structural model evaluation and modification: an interval estimation approach. Multivariate Behavioral Research 25, 173–180. doi:10.1207/s15327906mbr2502_4.

Straub, J., 2003. Psychologie und die Kulturen in einer globalisierten Welt, in: Kulturvergleichende Psychologie. Hogrefe, Verlag für Psychologie Göttingen, 543–566.

Thran, M., Boehnke, L., 2015. The value-based nationalism of Pegida. Journal for Deradicalization , 178–209.

Tomas, D., 1996. Transcultural space and transcultural beings. Boulder, Company: Westview Press.

Triandis, H., 1994. Culture and social behavior. McGraw-Hill series in social psychology, McGraw-Hill.

Tubadji, A., Möller, J., Nijkamp, P., 2014. Introduction to cultural research approaches: applications to culture and labour analysis. International Journal of Manpower 35, 2–10. doi:10.1108/ijm-08-2013-0191.

Van de Vijver, F.J.R., 2011. Capturing bias in structural equation modeling, in: Davidov et al. (2011). 3–34.

Van Deth, J.W., 1998. Equivalence in comparative political research, in: Comparative Politics: the problem of equivalence. Routledge. Volume 6, 1–19.

Wang, F., Wall, M.M., 2001. Modeling multivariate data with a common spatial factor. Technical Report. Division of Biostatistics, University of Minnesota.

Wang, F., Wall, M.M., 2003. Generalized common spatial factor model. Biostatistics 4, 569–582. doi:10.1093/biostatistics/4.4.569.

Wei, W., Joseph, K., Liu, H., Carley, K.M., 2016. Exploring characteristics of suspended users and network stability on Twitter. Social Network Analysis and Mining 6, 51. doi:10.1007/s13278-016-0358-5.

Wu, A.D., Li, Z., Zumbo, B.D., 2007. Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. Practical Assessment, Research and Evaluation 12, 1–26.