Universität Ulm
Institut für Stochastik

Statistical analysis and stochastic modeling of meteorological and paleogeographical space-time data

Dissertation

zur Erlangung des Doktorgrades Dr. rer. nat. der Fakultät für Mathematik und Wirtschaftswissenschaften der Universität Ulm

vorgelegt von

Björn Kriesche

aus Stollberg/Erzgeb.

2017

Amtierender Dekan:	Prof. Dr. Alexander Lindner
1. Gutachter:	Prof. Dr. Volker Schmidt
2. Gutachter:	Prof. Dr. Viktor Beneš
Tag der Promotion:	13.12.2017

Contents

Part I.

Preliminaries

1. Introduction

1.1. Motivation

During the past decades, the amount of (real or artificial) data that is collected or simulated in almost all fields of natural, technical or social sciences has grown exponentially, see, e.g., Turner et al. (2014). This is mostly driven by huge technical advances (greatly improved computer power, inexpensive and large hard drives, social media platforms, increased availability of image and video data, new innovative sensor technology, etc.) which enable a quick and cheap compilation of large and complex databases as well as extensive, high-resolution computer simulations. At the same time, new concepts and methods are constantly being developed, which allow for a more efficient acquisition, storage, processing, and analysis of such datasets. This trend encourages an intensified use of advanced stochastic models, statistical methods, and algorithms for Monte Carlo simulation to provide new insights to questions in all kinds of research areas. For example, in recent years stochastic models and methods have been successfully developed and applied at the Institute of Stochastics at Ulm University for a multitude of challenging problems covering such various fields as biology (e.g., Meinhardt et al., 2012; Lück et al., 2013), chemistry and physics (e.g., Brereton et al., 2014; Stenzel et al., 2014), climatology and meteorology (e.g., Rumpf et al., 2009; Elsner et al., 2013), dialectology (e.g., Rumpf et al., 2010), materials science (e.g., Gaiselmann et al., 2014; Neumann et al., 2016; Spettl et al., 2016; Handl et al., 2017; Westhoff et al., 2017), risk modeling (e.g., Christiansen et al., 2014; Kriesche et al., 2014), and telecommunication networks (e.g., Hirsch et al., 2015; Neuhäuser et al., 2016).

In this thesis, we consider two practical problems (one from meteorology and one combining questions from archeology and paleoecology), which we attempt to address by using stochastic models, methods, and simulation algorithms. Both applications have in common that all information in the underlying data are referenced by geographical coordinates and, additionally, have a temporal component. In fact, those are the most important information in the data as both problems considered in the following are related to the analysis and estimation of spatio-temporal distributions or the modeling and simulation of objects in space and time. The presence of space-time data suggests an extensive use of models and methods from disciplines that are referred to as stochastic geometry, spatial statistics, and geostatistics in literature but tools from other fields of applied mathematics will be considered as well (see Chapter 3). The main objective of this thesis is the development of new and innovative models and methods by combining existing spatial and spatio-temporal modeling approaches, statistical estimators, simulation algorithms, and optimization procedures from literature. In particular, we put a focus on the specific requirements the models and methods must meet in order to provide suitable solutions for the considered applications. This involves, e.g., that the developed approaches take into account special features in the available data, perform reasonably well with respect to observation data or independent results from literature and are implemented efficiently to allow short run times, which is crucial for an operational application.

The present thesis is embedded into two research collaborations between the Institute of Stochastics at Ulm University and several external partners, see Section 1.3. The results on stochastic models in weather prediction described in Part II of this thesis were achieved during a long-standing collaboration with the Deutscher Wetterdienst (DWD). Founded in 1952 as the main meteorological service of Germany, the DWD is responsible for providing timely, accurate and reliable weather forecasts and for monitoring meteorological and climatological conditions over Germany. A particularly challenging task is the issuing of weather warnings since severe weather events such as heavy precipitation, strong wind gusts, earth frost or hailstorms can cause both personal injury and high material damage. Thus, key customers of DWD range from civil protection, federal and regional authorities, provincial administrations of road construction, winter services, municipalities, business companies and media up to private customers and the general public. Classically, deterministic weather forecasts are computed based on numerical models describing the atmosphere, which is referred to as numerical weather prediction (NWP) in literature. Probabilistic forecasts can be derived using ensembles of numerical forecasts but it is generally acknowledged that the resulting predictions are subject to systematic errors. Therefore, the application of a wide range of probabilistic postprocessing methods has become of growing interest during the last decades to improve forecast quality. This involves that biased forecasts are corrected using data from meteorological observation systems such as rain gauges or anemometers. These systems typically represent measurements at fixed geographical locations, which is why the resulting calibrated probabilistic forecasts are related to single geographical locations (i.e., points, in mathematical terms), too. Consequently, DWD usually provides probabilities for the occurrence of weather events at given locations, which are denoted as point probabilities in this thesis. Especially for the issuing of weather warnings the consideration of point probabilities is often not sufficient as a critical situation can already arise if a (severe) weather event occurs somewhere in a region (or area, in mathematical terms) rather than at a fixed point. Two examples are given by the area of responsibility of a fire department, which is alarmed when there is heavy precipitation somewhere within that area, or by some warning area of DWD, for which a warning of freezing streets is issued in winter if there is precipitation of any small amount somewhere within that area (in combination with negative temperatures). A probability for a weather event occurring somewhere in an area is denoted as an area probability in this thesis. Accordingly, area probabilities are quantitatively different from point probabilities, e.g., an area probability of some weather event is always greater than or equal to a point probability of the same event at each fixed location within that area.

While a variety of methods for the derivation of precise and reliable point forecasts have been proposed in the literature, no general relationship is known for the analytical computation of area probabilities based on point probabilities (and their spatial correlations, which are expected to contain information about the spatial scale of the considered weather event). Existing approaches for probabilities of precipitation (Epstein, 1966; Krzysztofowicz, 1998) rely on restrictive assumptions, which make them inappropriate for the use in operational weather prediction. On the one hand, circular precipitation cells and uniformly distributed cell centers imply that point probabilities are equal for all locations in the considered forecast area, which does not allow for applications on a non-local scale such as the territory of Germany. On the other hand, additional information about the size of precipitation cells need to be provided by the forecaster, which prevents an automated generation of weather warnings, where an algorithmic computation of area probabilities based solely on point probabilities is desired. As a promising alternative, we suggest to compute or estimate area probabilities based on spatial stochastic models. In the present thesis, we consider the occurrence of precipitation exceeding a threshold $u \geq 0$ in mm (which includes the occurrence of precipitation of any amount for $u = 0$) and the occurrence of thunderstorms. While a variety of approaches to the spatial, temporal or spatio-temporal stochastic modeling of precipitation cells and precipitation amounts can be found in literature, the modeling of thunderstorms has not been considered yet. Unfortunately, certain limitations such as spatial and temporal stationarity, model fitting based on observation data or the independence of precipitation cells and precipitation amounts prevent an application of the models from literature in operational weather prediction. Therefore, one major goal of this thesis is the development of more robust and less restrictive stochastic models for precipitation cells, precipitation amounts and thunderstorm cells, which are suitable for the flexible, algorithmic computation of reliable area probabilities in a general context.

The results presented in Part III of this thesis were obtained in the framework of a research collaboration with the University of Ottawa and pertain to the analysis of relationships between the spatio-temporal distributions of human population and vegetation composition in North America during the Holocene (the geological epoch that began about 12000 years ago and lasts until today). While the rough distribution of different plant taxa in prehistoric North America has already been estimated on a continental scale (e.g., Williams et al., 2004), this has not yet been done for the distribution of human population. For a long time, there has been the prevailing opinion that the North American continent was settled via a land bridge connecting Alaska with eastern Siberia, see, e.g., Hoffecker et al. (1993). When a corridor opened between the Cordilleran and Laurentide Ice Sheets in western Canada about 14000 years ago, Native Americans migrated into all regions of North America that were not covered by ice (Bonatto and Salzano, 1997). However, today a large number of archaeologists consider it possible that a coastal migration took place by seafaring peoples from Asia (e.g., Goebel et al., 2008; Pedersen et al., 2016) or even other continents such as Australia (e.g., Skoglund and Reich, 2016), which would allow people to quickly reach all parts of the Americas before the ice-free corridor opened. In any case, archaeological findings (which are dated using the radiocarbon method) document that around 13000 years ago, Native American populations could be found in many different regions across the continent. The spatial patterns of subsequent population growth, however, are mostly unclear. Native Americans did not leave any written legacy, which is why no reliable information on population numbers and distributions is available until the arrival of Europeans about 500 years ago. In the present thesis, we thus attempt to statistically reconstruct the spatio-temporal distribution of populations in North America during the Holocene based on a large database of radiocarbon dates. Furthermore, using similar statistical techniques we also provide updated estimates of prehistoric vegetation abundances for a series of plant taxa based on fossil pollen data.

Another question of interest is how Native Americans interacted with their environment. Economic and technological activity of human populations often has a large effect on their natural environment and, vice versa, fluctuations in environmental conditions can have a significant influence on societies. For example, during the colonial period in North America, European settlers had an enormous impact on the landscape by deforesting large parts of eastern North America, although much has since regrown (Williams, 1989). Determining relationships between Native American populations and their environment before historical times, however, is much more complicated and can be done using (sub)fossil data. Two different points of view have been proposed in the literature to describe the nature of interactions between populations and their environment in prehistoric North America, see, e.g., Denevan (1992) or Vale (2002). One perspective commonly held through the 19th and 20th century suggests that in the pre-Columbian era North America was a ‘pristine landscape’ with low population densities, where vegetation was mostly unaltered by human activities. This view implies that there was no systematic impact of native populations on the environment and the primary factor causing changes in population numbers as well as in cultural or technological progress would have been environmental and climatic changes. However, an opposing viewpoint suggests that in large parts of North America, native populations altered the forests and plains through extensive land use, e.g., by burning down forests to allow for more intensive agriculture. Local and regional studies for North America have highlighted possible relationships between environmental changes and population sizes, see Delcourt and Delcourt (2004), Munoz and Gajewski (2010), Munoz et al. (2010) or Kelly et al. (2013). For example, in Munoz and Gajewski (2010) it is demonstrated that the introduction of agriculture at its northernmost range in Ontario led to a changing composition of forests over a period of several hundred years. If this was widespread, then it is entirely possible that human activity affected vegetation composition to a large extent across the continent. Therefore, we attempt to understand this interaction at continental scales by analyzing correlations between estimates of vegetation composition and population activity over the course of the past 13000 years.

1.2. Outline

This thesis is subdivided into three major parts. In Part I, the preliminaries, we provide a basis for the stochastic models and methods that are developed throughout this thesis. After the present introductory chapter we first give an overview of modern weather prediction as applied by DWD in Chapter 2 to allow for a better understanding of the available data and to provide a general context for the stochastic models that are developed later on in Part II. This includes an introduction to operationally applied atmospheric and surface observation systems, a brief discussion of NWP models for deterministic forecasting and a description of several approaches to probabilistic weather prediction (PWP) such as ensemble prediction and statistical postprocessing. Furthermore, we give a review of existing literature on the computation of area probabilities and on stochastic models for precipitation cells and precipitation amounts. In Chapter 3, we introduce the mathematical foundations that allow for a precise formal description of the developed models, methods, and simulation algorithms. Besides some general notation, we focus on tools and concepts of geostatistics (e.g., estimation of dependency structures in and between random fields), models and methods from stochastic geometry (such as random point processes and germ-grain models), and the application of nonparametric kernel methods in density estimation and regression analysis.

In Part II, we discuss several approaches to the spatial and spatio-temporal modeling of weather events with the purpose of estimating area probabilities. At first, Chapter 4 introduces a spatial stochastic model for the occurrence of precipitation, which can be used to compute area probabilities for arbitrary areas of interest based on point probabilities provided by DWD. In this context, precipitation cells are modeled as circular discs with random centers and a joint random radius, and methods are derived to algorithmically determine model characteristics without further input of the forecaster. This approach is then extended in Chapter 5 to the modeling of precipitation amounts. A randomly scaled response function is attached to each precipitation cell and the individual response functions are summed up to obtain precipitation amounts. Again, methods for the algorithmic computation of model characteristics are provided and it is described how the model can be used for the estimation of area probabilities for the occurrence of precipitation exceeding an arbitrary threshold. Finally, Chapter 6 describes a generalization of the proposed model of precipitation cells to the representation of thunderstorm cells using cluster processes. In addition to forecasted point probabilities of DWD we also take recent thunderstorm records into account, which allows to significantly increase the precision of predicted thunderstorm events for lead times of few hours ahead. All three models discussed in Part II are evaluated by comparing obtained area probabilities to high-resolution observation data.

Part III of the present thesis deals with the statistical analysis of archaeological and paleoecological data for North America. At first, Chapter 7 describes an approach to the statistical estimation of population intensity maps for the past 13000 years based on a comprehensive database of radiocarbon-dated archaeological material. For that purpose we propose a nonparametric method which accounts for several potential biases such as inhomogeneous sampling strategies, taphonomic loss or boundary effects. Furthermore, we give a brief interpretation of obtained results and perform a sensitivity and robustness analysis. Similar statistical techniques are applied in Chapter 8 to determine spatio-temporal estimates of vegetation intensity maps for a series of plant taxa that are major constituents of the forests and prairies of North America. As underlying data we use fossil pollen samples from a paleoenvironmental database, which can be considered as a quantitative index of past plant abundance. Once estimates of population and vegetation intensities are available, we perform a correlation analysis to find systematic relationships between demographic changes and environmental conditions in prehistoric North America. For that purpose, Chapter 9 discusses a statistical methodology to estimate spatio-temporal cross-correlation functions of vegetation and population intensities as well as cross-correlations of changes in vegetation and population at various temporal lags.

To conclude the thesis, Chapter 10 provides a summary of the main results and illustrates the insights that are obtained from the applied statistical methodology. Furthermore, we briefly discuss some open questions and suggest potential topics for future research.

1.3. Research collaborations

As mentioned in Section 1.1, the results described in the present thesis were obtained in several interdisciplinary research collaborations between the Institute of Stochastics at Ulm University and various external partners. The approaches presented in Part

II of this thesis for the stochastic modeling of weather events with the purpose of estimating area probabilities were developed and evaluated in a joint research project with the Research and Development of DWD. Furthermore, the spatial stochastic model introduced in Chapter 5 for the representation of precipitation amounts and the statistical procedures for the computation of model characteristics were established in collaboration with the Department of Probability and Mathematical Statistics at Charles University in Prague.

The statistical methods for the estimation of spatial population and vegetation intensity maps as well as the correlation analysis of obtained maps discussed in Part III were developed jointly with the Laboratory for Paleoclimatology and Climatology and the Department of Mathematics and Statistics at the University of Ottawa. Additionally, the Canadian Museum of History in Ottawa and the Department of Anthropology at the University of British Columbia in Vancouver were involved in the interpretation of population intensity maps provided in Section 7.

1.4. Software

An important aspect in developing the stochastic models and methods presented in this thesis is an efficient implementation to allow for an automatic operational application. Most concepts introduced in the following were implemented in Java or R, where we also rely on existing packages and libraries. On the one hand, we use several classes from the `GeoStoch`, a Java library which was jointly developed by the Institute of Stochastics and the former Institute of Applied Information Processing at Ulm University, see Mayer et al. (2004), and which is still extended regularly. In particular, we use classes for the handling of two-dimensional geometric objects (such as point patterns, discs, and convex polygons), the simulation of random point processes, the computation of distances between geographical coordinates, and the fitting of semivariogram models, where the classes described in Faulkner (2002) are used. In addition to the `GeoStoch` we incorporate classes from the `Parallel Java` library developed by the Department of Computer Science at Rochester Institute of Technology, see Kaminsky (2007), the `Apache Commons Math` library of the Apache Software foundation, see `http://commons.apache.org/proper/commons-math`, and the `JOptimizer` library for mathematical optimization problems, see `http://www.joptimizer.com`.

The statistical software package R is mainly used for visualization purposes. In particular, we import several packages that are specifically designed to facilitate the reading, processing and visualization of geographically referenced data (`maps`, `maptools`, `rgdal` and `geosphere`), the handling of geometric objects (`rgeos`) or the illustration of three-dimensional functions (`rgl`). Furthermore, two packages for the automatic selection and processing of paleoecological data are incorporated (`neotoma` and `analogue`).

2. Fundamentals of weather prediction

In order to facilitate the description of the available data and the mathematical framework used later on in this thesis in the context of modeling precipitation and thunderstorms, the present chapter provides some fundamentals of modern weather prediction as applied by DWD. We describe two routine procedures operated by DWD to record precipitation rates and thunderstorms in the territory of Germany (Section 2.1) and give some insight into the basics of numerical and probabilistic weather prediction (Sections 2.2 and 2.3). A more detailed overview of the discussed concepts and systems can also be found on the website of DWD (`www.dwd.de`). Furthermore, this chapter contains a short review of previous literature on the topics considered in Part II of the present thesis. To be more precise, in Section 2.4 the notations of point and area probabilities are introduced and an analytical relationship between both types of probabilities is presented under simplified conditions. As such approaches turn out to be inapplicable in operational weather prediction, we alternatively suggest to derive area probabilities based on spatial stochastic models, which are calibrated using point forecasts. For that purpose, a brief review of stochastic models for precipitation proposed in literature is provided in Section 2.5.

2.1. Weather observations

Reliable and exact high-resolution observation data are fundamental for weather services to perform their tasks such as providing precise weather forecasts and monitoring climatological developments. For that purpose, a large number of different observation systems is operated by DWD. On the one hand, DWD maintains a monitoring network of weather stations for surface observations (some manned by professionals or volunteers, some running fully automatic) that record weather data at single, fixed locations and regular time points (a few even providing temporally continuous data). Measured meteorological quantities include air pressure, temperature, humidity, wind velocity, wind direction, amount of precipitation, and duration of sunshine. Recorded raw data are analyzed, preprocessed, and archived into data bases to be used for the calibration of atmospheric observations (see below) or as input for NWP (see Section 2.2) and statistical postprocessing in PWP (see Section 2.3.2).

In many meteorological applications weather data with a much higher spatial resolution are required, which can only be obtained through systematic observation of the atmosphere (i.e., through remote sensing). The most popular remote sensing technologies used to provide spatially inclusive and comprehensive data of different meteorological quantities are weather surveillance radar and geostationary weather satellites. As we only consider data that are (mainly) derived from radar in this thesis, we briefly summarize the functionality of the polarimetric Doppler C-Band radar systems as operated by DWD. Using a transmitter and an antenna, a radar system generates impulses of electromagnetic waves, which are radiated into the atmosphere (bundled in a fixed direction). The electromagnetic waves propagate and hit the particles in the atmosphere, where a small amount of the waves is reflected back to the radar system. This so-called radar echo is then amplified, analyzed, and digitized, and the strength of the received signal can be considered as an indicator for the intensity of precipitation. Based on the fixed direction and the length of the time period between radiating the electromagnetic wave and receiving the reflected signal, the exact location of the reflecting particles can be determined. Changing the direction of the antenna according to a coordinated radar-scan-strategy (which is consistent for all systems in the radar network) allows to continuously monitor the atmosphere with a high spatial resolution. The entire scan procedure is repeated every 5 minutes providing a high temporal resolution as well. Radar reflectivities can be further processed to obtain data on, e.g., precipitation amounts or thunderstorm cells.

In order to provide spatially inclusive and comprehensive records of precipitation amounts, the routine procedure RADOLAN (Radar-Online-Aneichung) is applied by DWD. The RADOLAN system is based on a regular 1 km $\times$ 1 km lattice, which consists of 900 $\times$ 900 locations in central Europe ranging from 2.1°E to 15.7°E and 46.6°N to 54.9°N. Point coordinates that refer to the RADOLAN lattice are denoted as RADOLAN coordinates throughout this thesis. RADOLAN uses radar reflectivities, which are derived by the 17 radar stations of the operational weather radar network of DWD, see Winterrath et al. (2012). As reflectivities only provide indirect information on observed precipitation amounts, an adjustment is needed. For every full hour, aggregated radar reflectivities are merged with surface observations of more than 1,200 rain gauges at conventional meteorological sites, combining the advantages of both measurement technologies. The currently best possible RADOLAN product provides hourly quantitative precipitation data for all points of the RADOLAN lattice that are located within the territory of Germany with precipitation amounts rounded to a multiple of 0.1 mm. An additional clutter filter for hydrological applications can be applied in order to remove spurious pixel-scale precipitation events, see Winterrath and Rosenow (2007).

The second source of weather observation data being used in this thesis are derived from the NowCastMIX system, which archives thunderstorm records in high spatial and temporal resolution and provides various forecasts for the purpose of issuing short-term thunderstorm warnings. NowCastMIX combines thunderstorm records that are obtained every five minutes from three different approaches: CellMOS, KONRAD (Konvektionsentwicklung in Radarprodukten) and lightning sensors. CellMOS and KONRAD detect centers of thunderstorm cells based on (unadjusted) radar reflectivities, see Lang (2001). In order to do this, areas of adjacent pixels with radar reflectivities exceeding a certain threshold are determined and their barycenters are identified as centers of thunderstorm cells (large areas are subdivided into several individual cells). In addition, different cell characteristics are analyzed and recorded. Applications of data from KONRAD are described, e.g., in Wapler et al. (2012) and Wapler (2017). Moreover, CellMOS also provides probabilistic forecasts of thunderstorm events by combining detected cells with data from NWP, which are, however, not used in this thesis. As a third source of data, NowCastMIX interprets the coordinates of lightning strikes detected by lightning sensors as thunderstorm cell centers as long as they are not located within a radius of 10 km to a previously recorded cell center (lightning data are called LINET and are provided by nowcast GmbH). In NowCastMIX, thunderstorm cell centers are described using RADOLAN coordinates (see above) and single cells are modeled as discs with a fixed radius of 10 km. Additionally, several thunderstorm characteristics such as movement speed, movement direction and hail flag (an indicator for thunderstorm strength) are derived from radar reflectivities, using several radar processing methods together with lightning density and NWP data. Based on the movement speed and direction a warning cone with a propagation angle of 7.5° is computed for each thunderstorm cell describing the possible movement of the cell during the subsequent 60 minutes.

2.2. Numerical weather prediction

Generally speaking, weather forecasting involves the application of scientific methods to predict future states of the atmosphere using (surface and atmospheric) observations from past and present time periods. Almost all modern weather forecasts are based on numerical models describing the principles of atmospheric physics. We briefly sketch the basic functionality of NWP models, for more details see, e.g., Coiffier (2011) and Inness and Dorling (2013). In a numerical model the dynamic physical processes in the atmosphere are described using a system of differential equations and physical parameterizations. Based on Navier-Stokes equations from fluid dynamics, see Chorin and Marsden (1993), the spatio-temporal evolution of the most important meteorological variables such as temperature, wind velocity, air pressure, and water vapor are modeled. These complex differential equations can only be solved numerically, which requires a suitable discretization. For that purpose, the system of equations is transfered into a grid point model, which means that the spatio-temporal evolution of the considered meteorological variables is computed on a three-dimensional spatial grid covering the atmosphere from the surface up to a certain maximal height. Some NWP models are based on a global (i.e, world-wide) grid as for forecast ranges of five days and more the future weather at a fixed location (e.g., in Germany) can depend on the current state of the atmosphere in many regions of the earth. One of the most important characteristics of the model is the grid spacing, i.e., the horizontal distance of neighboring grid points. When choosing a possible grid spacing a compromise needs to be found between providing a sufficiently detailed representation of atmospheric structures and allowing the system of equations to be solved in a reasonable computation time. However, many physical phenomena in the atmosphere correspond to spatial scales that are smaller than the grid spacing and can thus not be resolved explicitly by the NWP models. These so-called sub-gridscale processes require additional parameterization schemes in order to describe their impact on the overall evolution of the atmosphere. Important sub-gridscale features include interaction of the atmosphere with solar and thermal radiation, cloud microphysics, convection and sub-gridscale orography. Altogether discretization and parameterization lead to a huge system of differential equations in time (typically 10^8 to 10^9 equations). Finally, numerical difference methods for initial value problems are applied to determine approximate solutions of the considered system in time steps ranging from several seconds to few minutes (depending on the chosen grid point model). For example, at DWD the Runge-Kutta method of third stage and the predictor-corrector method are used for that purpose, see, e.g., Butcher (2016). In order to apply such methods efficiently to the large system of equations, powerful supercomputers need to be available.

DWD operates two NWP models, which are briefly described in the following. As one of fourteen weather services in the world, DWD runs a global model called the ICON (Icosahedral Nonhydrostatic) model, see Zängl et al. (2015). It was introduced in 2015 and superseded the GME (Globalmodell Europa), which started operational work in 1999 as the first NWP model in the world using an icosahedral grid, see Majewski (1998) and Majewski et al. (2002). In the current version, the ICON model contains 2,949,120 horizontal equilateral triangles, whose centers form the points of the global ICON grid with an effective grid spacing of about 13 km. Furthermore, the model contains 90 vertical layers ranging from the surface to a height of about 75 km resulting in a total of more than 265 million grid points. The most important meteorological variables included in the ICON model are air density, virtual potential temperature (indicating air pressure), horizontal and vertical wind velocity, humidity, cloud water, cloud ice, rain, and snow. In operational use the ICON model first starts a data assimilation cycle in order to determine an optimal initial state for the forecast runs. A first guess, which is derived from a previous forecast, is combined with recent observations using a 3D variational assimilation method. Forecast runs are then performed every six hours:

at 00 UTC (Universal Time, Coordinated) and 12 UTC of each day with forecast ranges of up to 180 hours and at 06 UTC and 18 UTC for ranges up to 120 hours. Forecasts are written for each one-hour period up to a range of 78 hours and for every third one-hour period afterwards.

Primarily, DWD is responsible for providing forecasts and weather warnings for the territory of Germany. Thus, a second NWP model, the COSMO-DE, is used operationally, which has a much finer spatial resolution and thus produces more precise weather forecasts, see Baldauf et al. (2011). The COSMO-DE covers Germany, Switzerland, Austria, and parts of neighboring countries by a regular grid consisting of 421×461 points with a grid spacing of 2.8 km and 50 vertical layers, i.e., a total of 9.7 million grid points. This extremely high horizontal resolution enables a more realistic representation of topographic features and allows to better resolve deep convection such as cumulonimbus clouds, providing much more precise forecasts of heavy precipitation, thunderstorms, hail, and strong wind gusts. Accordingly, the high resolution also decreases the need for additional parameterization. Since small-scale meteorological objects such as cumulonimbus clouds typically have shorter life times than large-scale high and low pressure systems, COSMO-DE provides forecasts every three hours for forecast ranges up to 27 hours only. The COSMO-DE is not operated globally, making the use of boundary values from the ICON model necessary. Furthermore, radar data are taken into account as initial values by using a method called 'Latent Heat Nudging'.

2.3. Probabilistic weather prediction

NWP models as described in Section 2.2 provide deterministic forecasts, i.e., they compute one single future state of the atmosphere for any considered point in time. In applications, however, certain approximations, discretizations, and simplifications concerning model equations and implementation are made and measurement errors in observation data (which are used as initial values) often occur. Thus, deterministic forecasts from NWP models are subject to uncertainties, which can be taken into account by providing probabilistic forecasts.

2.3.1. Ensemble prediction

The most popular method to account for uncertainties in NWP models is using ensemble prediction systems. An ensemble consists of different forecast scenarios, called ensemble members, that are computed based on slightly different (but realistic) initial values, configurations, and parameterizations of the NWP model. This variation has an effect on the obtained deterministic forecasts, which gets more remarkable the longer the forecast range is (due to the chaotic nature of the atmosphere). Therefore, an ensemble can be interpreted as a sample of equally likely future weather scenarios. If the number of ensemble members is sufficiently large, the ensemble can be used to statistically estimate probabilistic characteristics such as expectations and variances of meteorological variables (e.g., temperature, precipitation amount, wind velocity) or probabilities for the occurrence of certain weather events (e.g., precipitation, thunderstorms, wind gusts).

At DWD the ensemble prediction system COSMO-DE-EPS has been used operationally since the year 2012, see Gebhardt et al. (2011). The COSMO-DE-EPS computes 20 ensemble members based on the COSMO-DE model, see Section 2.2, which are obtained by using different boundary conditions, initial values, and configurations of physical parameterization as well as perturbations in the soil moisture. Ensemble prediction has a particular high relevance for the COSMO-DE model since small-scale processes (which can be represented by the COSMO-DE due to its fine spatial resolution) are subject to a higher degree of uncertainty than forecasts of processes on a larger scale such as low and high pressure systems. A similar ensemble prediction system for the ICON model is currently under development at DWD and is planned to be operationally available by the end of 2017.

2.3.2. Statistical postprocessing

It is generally acknowledged that forecasts from NWP models and ensemble prediction systems are subject to systematic errors. Especially forecasts at grid points near the surface tend to be underdispersed, see Gebhardt et al. (2011). In order to further increase forecast quality, to correct biases, and to forecast additional meteorological variables that are not directly included in the NWP models or the ensemble prediction systems, a statistical postprocessing is performed. For that purpose, DWD uses Model Output Statistics (MOS), see, e.g., Glahn and Lowry (1972), Knüpffer (1996) or Wilks (2011). MOS involves a statistical optimization of forecasts by comparing long historical time series of observed meteorological variables with the corresponding uncalibrated predictions provided by NWP models and ensemble prediction systems. In a development cycle, statistical relationships and systematic differences between (uncalibrated) forecasts and observation data are analyzed and modeled using stepwise regression equations. For each predictand (i.e., each meteorological variable to be forecasted) the most relevant predictors are selected from a set of about 300 variables as long as they are statistically significant. Most predictors are derived from the numerical models and ensemble prediction systems but also recent observations are used (in particular for short range forecasts) to provide meteorological persistence. Then, for continuous variables, such as precipitation amount, temperature or wind velocity, linear regression is applied, whereas logistic regression is used for probabilistic predictands such as probabilities for the occurrence of certain weather events. In operational use the fitted regression models are then applied to derive forecasts based on the current output of the NWP models and ensemble prediction systems and recent observations. Furthermore, MOS is also used operationally to combine numerical or probabilistic weather predictions from different forecast systems.

We briefly present two MOS products of DWD that provide calibrated probabilities as data input for the stochastic models developed in Part II of this thesis (note that a variety of other MOS systems and products is operated by DWD for different applications). At first the so-called MOSMIX system is considered. MOSMIX calibrates and combines deterministic forecasts of two numerical models: the ICON model, see Section 2.2, and the IFS (Integrated Forecasting System) of the European Centre for Medium-Range Weather Forecasts (ECMWF) in Reading. Forecasts are provided four times a day with forecast ranges of up to ten days for a network of 4,000 weather stations worldwide (most of them being located in western and central Europe). Furthermore, MOSMIX forecasts include about 150 meteorological variables such as expected temperature, dew point temperature, wind velocity and direction, precipitation amount, air pressure, visibility, and sun shine duration as well as a large number of probabilities for the occurrence of several weather events. A second operational system of DWD called ModelMIX even allows to provide calibrated forecasts at arbitrary locations in Germany. For that purpose the territory of Germany is subdivided into nine climatic zones and a joint MOS regression is performed for all stations in each zone separately. The obtained fitted regression models can then be applied at the original sites of the weather stations or at any other location within the corresponding climatic zone. In the latter case, observations need to be interpolated to the specified location but the numerical models and ensemble prediction systems are always evaluated for the exact location (i.e., the next grid point). Currently, a statistical postprocessing as described above is performed separately for forecasts from the ICON model, the IFS of ECMWF and the ensemble prediction system COSMO-DE-EPS, see Section 2.3.1. ModelMIX then combines the individual results (again using a MOS technique) in order to obtain an statistically optimal probabilistic weather forecast.

2.4. Point and area probabilities

The importance of probabilistic methods in operational weather prediction has grown steadily during the last decades. The present practice of most weather services such as DWD is to calibrate forecasts with synoptic meteorological observation systems (e.g., rain gauges at weather stations, see Section 2.1) that provide measurements at given geographical locations. Accordingly, obtained probabilistic forecasts represent the chances of weather events to occur at fixed locations (or points) and are thus denoted as point probabilities in the following. However, in many applications, particularly when providing weather warnings, the consideration of point probabilities is not sufficient, e.g., when a critical situation arises if the weather event occurs somewhere in an area (rather than at a fixed point). Thus, there is a strong interest in the computation of reliable probabilities for the occurrence of weather events somewhere in a given region (or area), which are denoted as area probabilities in this thesis. Until recently, no operationally applicable methods to the computation of unbiased area probabilities were known. One example is given by the ensemble prediction system COSMO-DE-EPS of DWD, see Section 2.3.1, which also provides uncalibrated area probabilities for quadratic areas with a size of 28 km $\times$ 28 km. However, since these probabilities are directly derived from the numerical COSMO-DE model (without any further post-processing), they contain a significant bias, see the forecast verification performed in Hess et al. (2017).

Relationships between point and area probabilities under general conditions are still unknown. Obviously, an area probability of some weather event is always larger than or equal to a point probability of the same event for any fixed location within that area. The difference of both kind of probabilities, however, depends on many factors, e.g., the size of the considered area or the scale and the dependence structure of the weather event. Thus, representation formulas for the computation of area probabilities based on point probabilities can only be derived under simplified conditions. In Epstein (1966) and Krzysztofowicz (1998) the authors consider the occurrence of precipitation, where they assume circular precipitation cells with a globally fixed radius that are uniformly distributed over the considered domain. In this particular case, the area probability $\pi(B)$ of any circular forecast area B (whose distance to the boundary of the considered domain is not smaller than the radius of the precipitation cells) is given by

$$\pi(B) = 1 - (1 - p(s))^{(1+Q^{-\frac{1}{2}})^2} \,,$$

where $p(s)$ denotes the point probability at some arbitrary location $s \in B$ and $Q = C/A$ is called a cell/area ratio, with $C > 0$ being the area covered by a precipitation cell and $A = \nu_2(B) > 0$ (where ν_2 denotes the 2-dimensional Lebesgue-measure, see Section 3.1.1). In this setting, we always get that $\pi(B) > p(s)$ whenever $p(s) \in (0,1)$ and that $\pi(B) \to p(s)$ for $Q \to \infty$. In other words, the difference between point and area probabilities gets less significant when the forecast area B is small compared to the area C of a precipitation cell. In Krzysztofowicz (1998) additional representation formulas for the expectation and variance of the fraction of B that is covered by precipitation based on $p(s)$ and $\pi(B)$ are derived. However, the very restrictive assumptions mentioned above make such formulas inappropriate for the automated generation of weather forecasts. Model parameters such as the radius of precipitation cells have to be determined by the forecaster and missing spatial non-stationarity prevents an application on a non-local scale such as the territory of Germany.

2.5. Review of stochastic models for precipitation

As an alternative to the use of direct computation formulas, area probabilities can be estimated using stochastic models. Since the 1980s there has been an extensive literature on the stochastic modeling of precipitation cells and precipitation amounts. Several approaches to the temporal representation of precipitation at single and multiple sites have been proposed, which are based on, e.g., Poisson cluster processes, in particular Neyman-Scott and Bartlett-Lewis rectangular pulse models and their extensions (Rodriguez-Iturbe et al., 1987, 1988; Chandler, 1997; Cowpertwait et al., 2007, 2011; Kaczmarska et al., 2014), Markov processes (Woolhiser and Osborn, 1985; Hughes et al., 1999; Srikanthan and Pegram, 2009), doubly stochastic Poisson processes (Ramesh et al., 2013; Thayakaran and Ramesh, 2017) or generalized linear models (Yang et al., 2005).

Clearly, temporal models cannot be used for the estimation of area probabilities but they motivate more sophisticated concepts for the continuous spatial and spatio-temporal representation of rain events. Again, a majority of approaches suggested in literature relies on random point processes. In Cox and Isham (1988) centers of precipitation cells are modeled using stationary spatio-temporal Poisson point processes and single cells are represented as discs with random radii, random movement speeds, and random precipitation amounts (constant over each cell). Several probabilistic characteristics are computed based on the proposed model but neither model fitting nor empirical applications are considered. In Cowpertwait (1995) this approach is extended by using temporal cluster processes and by allowing for different cell types. Additionally, a brief empirical example using point data from rain gauges is given. Further enhancements for this type of model include the introduction of elliptical precipitation cells and the development of fitting methods (generalized method of moments, spectral methods) for the estimation of model parameters based on radar data, see Northrop (1998) and Wheater et al. (2000). Detailed summaries of the mentioned point-process-based modeling approaches (also for the single-site and multi-site case) can be found in Onof et al. (2000) and Wheater et al. (2005). The most commonly used point process model for precipitation is the STNSRP (spatio-temporal Neyman-Scott rectangular pulses) model (Burton et al., 2008), which can briefly be outlined as follows. At first, storm centers are modeled using a stationary spatio-temporal Poisson point process. If a storm occurs, then centers of precipitation cells within the storm are represented by another stationary spatial Poisson point process and each cell is attached with a random radius, a random lifetime, and a random precipitation amount (constant over the cell), which are all exponentially distributed. The precipitation amount at any point can be obtained by summing up the amounts of all individual precipitation cells covering that point. Finally, a location-dependent scaling factor is considered to account for orography effects on local precipitation amounts. The model parameters can be fitted based on rain gauge data using a complex numerical optimization scheme. In Burton et al. (2010) the

STNSRP model is extended by allowing precipitation cells to occur inhomogeneously in space to also account for the effects of orography to the local intensity of precipitation cells. A more realistic representation of precipitation amounts for single precipitation cells is proposed in Rodriguez-Iturbe et al. (1986), where stationary shot-noise fields are considered for the spatial modeling of precipitation. To be more precise, precipitation cells are represented as discs whose centers are modeled by Poisson or Neyman-Scott point processes and to each cell a randomly scaled response function is attached. The authors derive several theoretical characteristics of their model and make a comparison for different response functions (linear, quadratic or exponential). Model fitting based on observed data, however, is only described vaguely. There are also approaches to the continuous spatial and spatio-temporal modeling of precipitation suggested in the literature that are not solely based on random point processes. For example, Smith and Krajewski (1987) consider a combination of random point processes with Markov chains and in Lanza (2000) a conditional simulation algorithm for intermittent rain fields based on stationary Gaussian random fields is proposed. A more recent approach to the representation of high-resolution precipitation in space and time is given by the STREAP (space-time realizations of areal precipitation) model, see Paschalis et al. (2013) and Peleg et al. (2017), where precipitation fields are modeled in three steps. At first, storm arrivals are described by alternating temporal renewal processes. Then, the temporal evolution of mean areal statistics such as the fraction of wet area and the mean areal precipitation intensity are represented using bivariate Gaussian processes. Finally, the spatio-temporal evolution of the storm structure is modeled based on latent Gaussian random fields and autoregressive moving average time series. A validation has shown that the STREAP model is able to provide even more realistic realizations (concerning intermittency, spatial and temporal correlations, growth and decay of storms over time) than the STNSRP model.

Unfortunately, certain limitations prevent the use of the above-mentioned models in operational weather prediction. Most approaches assume temporal and spatial stationarity and model fitting procedures (mostly for large parameter sets) always rely on observation data from rain gauges or radar systems. In operational weather forecasting, however, it is crucial to account for the (permanently changing) weather conditions in the current period and future forecast periods rather than for weekly, monthly or yearly averages, which contradicts the assumption of temporal stationarity. Furthermore, forecasters are typically interested in stochastic models to be applied on a non-local scale, which does not allow for spatial stationarity assumptions due to different meteorological (areas of low or high pressure, changing over time) and geographical (plains or mountainous regions, not changing over time) conditions. Although effects of orography are taken into consideration in some approaches, see, e.g., Cowpertwait (1995), Burton et al. (2008) and Burton et al. (2010), none of the proposed models accounts for spatially varying meteorological conditions in fixed forecast periods. On the other hand, model fitting based on precipitation data from radar or rain gauges is not suitable, either. Observation data from periods prior to the forecast period become unrepresentative already after few hours and data for the (future) forecast period are not available at the time the forecast is made. Further limitations occurring in some of the mentioned approaches include constant precipitation amounts per precipitation cell, independence of precipitation cells and precipitation amounts or the absence of model fitting procedures. Although not directly applicable in operational weather prediction, the previous approaches to the spatial modeling of precipitation from literature provide a valuable basis for the models developed in Part II of the present thesis.

In contrast to precipitation there is hardly any literature for the stochastic modeling of thunderstorms available. One of the few examples is Li (2000), where a model for wind loads in thunderstorms is proposed at single sites but approaches to the spatially continuous modeling of thunderstorm cells do not seem to exist.

3. Mathematical foundations

In the present chapter, we discuss some mathematical basics that are essential for a rigorous description of the stochastic models and statistical methods considered in the subsequent chapters of this thesis. After giving some general definitions in Section 3.1, we address the following topics:

1. *random fields* with an emphasis on characterization and estimation of dependencies (Section 3.2),
2. models and methods of *stochastic geometry*, in particular (random) point processes and related germ-grain models (Section 3.3),
3. methods and estimators from *multivariate kernel smoothing* with a focus on their application in density estimation and regression analysis (Section 3.4).

3.1. Basic definitions

3.1.1. General notation

First of all, we introduce some general notation that is repeatedly used throughout this thesis. By $\mathbb{N} = \{1, 2, \ldots\}$ we denote the set of natural numbers, by $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ the set of nonnegative integers, and by $\mathbb{R}$ the set of real numbers. For any dimension $d \in \mathbb{N}$, let $\mathbb{R}^d$ be the set of d-dimensional vectors with real components (also denoted as d-dimensional Euclidean space), where $o = (0, \ldots, 0)^\top \in \mathbb{R}^d$ is the origin and $\| \cdot \|_d$ is the Euclidean norm on $\mathbb{R}^d$. We write $b(x, r) = \{y \in \mathbb{R}^d : \|x - y\|_d \leq r\}$ for the d-dimensional (closed) ball with center $x \in \mathbb{R}^d$ and radius $r > 0$. By $\mathcal{B}(\mathbb{R}^d)$ we denote the Borel σ-algebra on $\mathbb{R}^d$, by $\mathcal{B}_0(\mathbb{R}^d)$ the family of bounded Borel sets in $\mathbb{R}^d$, and by $\nu_d : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ the d-dimensional Lebesgue measure. For any $B \in \mathcal{B}_0(\mathbb{R}^d)$, let $\mathcal{B}(B)$ be the Borel σ-algebra that is generated by all open subsets of B.

For $m, n \in \mathbb{N}$, the set of all $m \times n$ matrices with real coefficients is denoted by $\mathbb{R}^{m \times n}$ and for any $A \in \mathbb{R}^{m \times n}$, the matrix $A^\top \in \mathbb{R}^{n \times m}$ is the transpose of A. We write $A = \mathrm{diag}(a_1, \ldots, a_m) \in \mathbb{R}^{m \times m}$ for the matrix with entries $a_1, \ldots, a_m \in \mathbb{R}$ on the main diagonal and zeros everywhere else and $I = \mathrm{diag}(1, \ldots, 1) \in \mathbb{R}^{m \times m}$ for the m-dimensional unit matrix. For $A \in \mathbb{R}^{m \times m}$ let A^{-1} be the inverse (provided that it exists), $\det(A)$ the determinant, and $\mathrm{tr}(A)$ the trace of A, the latter being defined as the sum of its diagonal elements.

For any subset B of $\mathbb{R}^d$, we write $\overline{B}$ for its topological closure, $\#B$ for its cardinality, and $\mathbb{1}_B : \mathbb{R}^d \to \{0,1\}$ for the indicator function of B, which is defined by

$$\mathbb{1}_B(x) = \begin{cases} 1 & \text{if } x \in B, \\ 0 & \text{if } x \notin B, \end{cases} \qquad x \in \mathbb{R}^d.$$

For any translation vector $x \in \mathbb{R}^d$ and any rotation $\delta : \mathbb{R}^d \to \mathbb{R}^d$ around the origin, we write $B + x = \{y + x, y \in B\}$ and $\delta B = \{\delta(y), y \in B\}$ for the translation of B by x and the rotation of B by δ, respectively. The Minkowski sum of two sets $B_1, B_2 \subset \mathbb{R}^d$ is defined by $B_1 \oplus B_2 = \{x + y, x \in B_1, y \in B_2\}$.

By $\mathrm{supp}(f) = \overline{\{x \in \mathbb{R}^d : f(x) \neq 0\}}$ we denote the support of any real-valued function $f : \mathbb{R}^d \to \mathbb{R}$. In addition, provided that f has continuous first- and second-order partial derivatives, $\nabla_f(x) \in \mathbb{R}^d$ denotes the gradient of f at $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, which contains all first-order partial derivatives $\frac{\partial}{\partial x_i} f(x)$ for $i = 1, \ldots, d$ and $\mathcal{H}_f(x) \in \mathbb{R}^{d \times d}$ denotes the Hessian matrix of f at $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, which consists of the second-order partial derivatives $\frac{\partial^2}{\partial x_i \partial x_j} f(x)$ for $i, j = 1, \ldots, d$.

3.1.2. Random elements

Probability theory is the branch of mathematics that addresses the modeling of random phenomena and experiments. In order to formally describe stochastic models (as well as statistical methods which are typically based on models), it is customary to first define a suitable underlying probability space. A *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ is a triplet consisting of

1. an arbitrary set $\Omega \neq \emptyset$, which is denoted as *sample space*,
2. a σ-algebra $\mathcal{F}$ on Ω, where each element $A \in \mathcal{F}$ is called an *event*, and
3. a probability measure $\mathbb{P}$, i.e., a σ-additive mapping $\mathbb{P} : \mathcal{F} \to [0,1]$ with $\mathbb{P}(\Omega) = 1$, which assigns a probability $\mathbb{P}(A)$ to each event $A \in \mathcal{F}$.

In the following, unless specified more precisely, we always consider an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

In probability theory, a large variety of stochastic objects is considered for modeling purposes. The most general notation to describe such objects is that of a random element.

Definition 3.1.1 (Random element)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(S, \mathcal{B})$ be a measurable space, i.e., $S \neq \emptyset$ is some nonempty set and $\mathcal{B}$ is a σ-algebra of subsets of S. A mapping $X : \Omega \to S$ is called a *random element* with values in $(S, \mathcal{B})$ if X is $(\mathcal{F}, \mathcal{B})$-measurable, i.e.,

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}, \quad B \in \mathcal{B}.$$

For each $\omega \in \Omega$ we call $X(\omega)$ a *realization* of X.

Example 3.1.1. Examples of random elements include random variables ($S = \mathbb{R}$, $\mathcal{B} = \mathcal{B}(\mathbb{R})$), random vectors ($S = \mathbb{R}^d$, $\mathcal{B} = \mathcal{B}(\mathbb{R}^d)$) with finite dimension $d \geq 2$, random functions (see Section 3.2.1), random measures (see Section 3.3.1), random point processes (see Section 3.3.2), and random closed sets (see Section 3.3.6).

Definition 3.1.2 (Distribution of a random element)**.** Let X be a random element defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a measurable space $(S, \mathcal{B})$. The *distribution* of X is the probability measure P_X defined on $(S, \mathcal{B})$ by

$$P_X(B) = \mathbb{P}(X \in B) = \mathbb{P}\big(X^{-1}(B)\big), \quad B \in \mathcal{B}.$$

For any random element Y defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(S, \mathcal{B})$, we write $X \overset{\mathrm{d}}{=} Y$ if Y has the same distribution as X, i.e., $P_X(B) = P_Y(B)$ for all $B \in \mathcal{B}$.

Definition 3.1.3 (Independent random elements)**.** Consider an arbitrary index set $\mathcal{T} \neq \emptyset$ and a family $\{X(t), t \in \mathcal{T}\}$ of random elements defined on a joint probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a measurable space $(S, \mathcal{B})$. The random elements $\{X(t), t \in \mathcal{T}\}$ are called *(mutually) independent* if for any finite subset $\{t_1, \ldots, t_n\} \subset \mathcal{T}$ with $n \in \mathbb{N}$ it holds that

$$\mathbb{P}(X(t_1) \in B_1, \ldots, X(t_n) \in B_n) = P_{X(t_1)}(B_1) \cdot \ldots \cdot P_{X(t_n)}(B_n), \quad B_1, \ldots, B_n \in \mathcal{B}.$$

3.2. Random fields

In many statistical applications data are considered that can be interpreted as functions depending on a time index, a spatial location or both. For the modeling of such data (but also for the description of more complex random objects) random functions are typically used in literature. In this section, the concept of random functions is introduced to the reader, where we mainly focus on random functions whose domain is a subset of the d-dimensional Euclidean space $\mathbb{R}^d$. A particular emphasis is put on characteristics describing the dependence structure of random fields and on their estimation in a geostatistical context.

3.2.1. Random functions

Random functions are, similar to random elements, very general and flexible stochastic objects. For example, the class of random functions does not only include random vectors, processes, and fields but also more sophisticated objects such as random (counting) measures, see Section 3.3. For an overview of the general theory of random functions we recommend Yaglom (1987a) and Yaglom (1987b), whereas the special case of random fields is discussed in, e.g., Ivanov and Leonenko (1989), Ramm (2005) or Adler (2010) (with a focus on estimation theory for random fields in the first two).

Definition 3.2.1 (Random function)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(S, \mathcal{B})$ an arbitrary measurable space with $S \neq \emptyset$, and $\mathcal{T} \neq \emptyset$ an arbitrary index set. A family $X = \{X(t), t \in \mathcal{T}\}$ of random elements with values in S is called a *random function* with index set $\mathcal{T}$ and state space S.
To be more precise, X is a mapping from the product space $\Omega \times \mathcal{T}$ to S such that $X(\cdot, t) : \Omega \to S$ is $(\mathcal{F}, \mathcal{B})$-measurable for each $t \in \mathcal{T}$, where we usually write $X(t)$ instead of $X(\cdot, t)$. For each $\omega \in \Omega$ we call $\{X(\omega, t), t \in \mathcal{T}\}$ a *realization* or a *trajectory* of X.

Example 3.2.1. The class of random functions includes a wide range of stochastic objects, which are known under different names in the literature (depending on the index set and the state space). Consider a random function $X = \{X(t), t \in \mathcal{T}\}$ with index set $\mathcal{T}$ and state space $S \subset \mathbb{R}$.

1. If $\mathcal{T} = \{1, \ldots, d\}$ for some $d \in \mathbb{N}$, then X is a d-dimensional *random vector*. In this case, we usually write $X = (X_1, \ldots, X_d)$ instead of $X = \{X(1), \ldots, X(d)\}$.

2. If $\mathcal{T}$ is a countable subset of $\mathbb{R}$, then X is called a *stochastic process in discrete time*. As its name implies, a stochastic process in discrete time is often assumed to describe the development of some random value over time (which is why the index set is denoted by $\mathcal{T}$). Some authors even use the term stochastic process as a synonym for random function, see, e.g., Kallenberg (2002). Popular examples of stochastic processes in discrete time are given by Markov chains with discrete state space S, see, e.g., Behrends (2000).

3. If $\mathcal{T}$ is an uncountable subset of $\mathbb{R}$, e.g., $\mathcal{T}$ is some finite or infinite interval, then X is called a *stochastic process in continuous time*. There is a wide range of different stochastic processes considered in the literature such as Markov processes, renewal processes, Gaussian processes, Lévy processes or martingales. For a general overview see, e.g., Grimmett and Stirzaker (2001) or Kallenberg (2002).

4. If $\mathcal{T}$ is a subset of the d-dimensional Euclidean space $\mathbb{R}^d$ for some $d \geq 2$, then X is denoted as *random field*. Random fields play an important role in spatial

statistics and geostatistics since they are particularly suitable for the spatially continuous modeling of geographically-referenced data.

5. If $\mathcal{T}$ is a subset of the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ on $\mathbb{R}^d$, then X is called a *set-indexed random function*. An important example is given by the class of random measures, which is formally introduced in Section 3.3.1.

One can show that a random function can not only be interpreted as a family of random elements but also as a single random element taking values in a more complex space. For any measurable space $(S, \mathcal{B})$ with $S \neq \emptyset$ and any index set $\mathcal{T} \neq \emptyset$, let $S^{\mathcal{T}}$ denote the set of all functions $f : \mathcal{T} \to S$ and $\mathcal{B}^{\mathcal{T}}$ the smallest σ-algebra of subsets of $S^{\mathcal{T}}$ that contains the sets $\{f \in S^{\mathcal{T}} : f(t) \in B\}$ for all $t \in \mathcal{T}$ and $B \in \mathcal{B}$.

Theorem 3.2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(S, \mathcal{B})$ a measurable space with $S \neq \emptyset$, and $\mathcal{T} \neq \emptyset$ an index set. Furthermore, consider the measurable space $(S^{\mathcal{T}}, \mathcal{B}^{\mathcal{T}})$ as introduced above. The mapping $X : \Omega \times \mathcal{T} \to S$ with $X(\omega, \cdot) = \{X(\omega, t), t \in \mathcal{T}\}$ for $\omega \in \Omega$ is a random function with index set $\mathcal{T}$ and state space S in the sense of Definition 3.2.1 if and only if X is a random element with values in $(S^{\mathcal{T}}, \mathcal{B}^{\mathcal{T}})$, i.e., if the mapping that assigns the function $X(\omega, \cdot) \in S^{\mathcal{T}}$ to each $\omega \in \Omega$ is $(\mathcal{F}, \mathcal{B}^{\mathcal{T}})$-measurable.*

Proof. See Kallenberg (2002), Lemma 3.1.

Since a random function $X = \{X(t), t \in \mathcal{T}\}$ can be considered as a random element, its distribution P_X is given according to Definition 3.1.2. However, in most cases it is impossible to find a closed representation of P_X. Thus, we introduce the notation of the finite-dimensional distributions of X, which uniquely characterize the distribution P_X.

Definition 3.2.2 (Finite-dimensional distributions)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a random function with index set $\mathcal{T}$ and state space S. For fixed $n \in \mathbb{N}$ and $t_1, \ldots, t_n \in \mathcal{T}$, consider the distribution $P_X^{(t_1,\ldots,t_n)}$ of the random vector $(X(t_1), \ldots, X(t_n))$, i.e.,

$$P_X^{(t_1,\ldots,t_n)}(B_1, \ldots, B_n) = \mathbb{P}(X(t_1) \in B_1, \ldots, X(t_n) \in B_n), \quad B_1, \ldots, B_n \in \mathcal{B}.$$

Then, the family $\{P_X^{(t_1,\ldots,t_n)}, n \in \mathbb{N}, t_1, \ldots, t_n \in \mathcal{T}\}$ is called the family of *finite-dimensional distributions* of X.

Theorem 3.2.2. *Let $X = \{X(t), t \in \mathcal{T}\}$ and $Y = \{Y(t), t \in \mathcal{T}\}$ be random functions with index set $\mathcal{T}$ and state space S. Then, $X \overset{d}{=} Y$ if and only if*

$$(X(t_1), \ldots, X(t_n)) \overset{d}{=} (Y(t_1), \ldots, Y(t_n)), \quad n \in \mathbb{N}, t_1, \ldots, t_n \in \mathcal{T}.$$

In other words, the random functions X and Y have the same distribution if and only if they have the same finite-dimensional distributions. This implies that the family of finite-dimensional distributions of a random function uniquely determines its distribution.

Proof. See Kallenberg (2002), Proposition 3.2.

3.2.2. Random fields in geostatistics

In this thesis, we mainly consider random fields with the purpose of analyzing and modeling geostatistical data. *Geostatistics* denotes a branch of statistics, which is, according to Montero et al. (2015), defined as "the study of regionalized phenomena, that is, phenomena that stretch across space and which have a certain spatial organization or structure." The field of geostatistics emerged in the mining industry in the early fifties (Wackernagel, 2003) and rapidly evolved since then finding applications in a myriad of disciplines such as geology, hydrology, meteorology, geography, forestry, environmental science, ecology, and agriculture. A defining feature of geostatistical data is that the investigated variables are indexed by geographical locations and are (theoretically) observable everywhere in a fixed, non-random domain $\mathcal{T} \subset \mathbb{R}^d$. In practice, however, the variables are only observed at a finite sequence of deterministic locations $t_1, \ldots, t_n \in \mathcal{T}$ with $n \in \mathbb{N}$. In most references, the observed variables $x(t_1), \ldots, x(t_n)$ at $t_1, \ldots, t_n$ are denoted as *regionalized values* to emphasize their dependence on the corresponding locations. The (unobserved) family $\{x(t), t \in \mathcal{T}\}$ describing the values of the investigated variable at all locations in the domain $\mathcal{T}$ is called a *regionalized variable*.

A common approach in geostatistics is to interpret a regionalized variable $\{x(t), t \in \mathcal{T}\}$ as realization of some real-valued random field $\{X(t), t \in \mathcal{T}\}$. Accordingly, we assume that the index set $\mathcal{T}$ is a closed subset of the Euclidean space $\mathbb{R}^d$, where $d \geq 2$ and $\nu_d(\mathcal{T}) > 0$, and that the state space S is a subset of $\mathbb{R}$. Furthermore, if $\mathcal{T}$ is bounded, we define

$$r_0 = \max\{r \geq 0 : \text{for each } r' \leq r \text{ there are } s, t \in \mathcal{T} \text{ such that } \|s - t\|_d = r'\}. \quad (3.1)$$

If $\mathcal{T}$ is unbounded, then we assume that for each $r \geq 0$ there are locations $s, t \in \mathcal{T}$ such that $\|s - t\|_d = r$. Unlike most references, we keep the notation $\mathcal{T}$ for the index set in order to maintain consistency with the theory of general random functions provided in Section 3.2.1 (in geostatistics the notation $\mathcal{D}$ is often used instead) but from now on, we refer to $\mathcal{T}$ as a domain rather than as an index set. Furthermore, in a geostatistical context, the family of finite-dimensional distributions of a random field is sometimes referred to as the field's *spatial distribution* (Chilès and Delfiner, 2012). When analyzing geostatistical data, the focus of interest often lies either on the dependence structure of the underlying random field or, in the case that regionalized values of more than one variable are given, on the dependence structure between the corresponding random fields. Popular characteristics quantifying such dependencies include covariance functions, cross-covariance functions, and semivariograms, which are discussed in the following. More details on geostatistics (for purely spatial data) are provided in Journel and Huijbregts (1978), Cressie (1993), Wackernagel (2003), Schabenberger and Gotway (2005), Diggle and Ribeiro Jr. (2007), and Chilès and Delfiner (2012). In more recent times, there has been a growing interest in models and methods for spatio-temporal datasets, see, e.g., Cressie and Wikle (2011) or Montero et al. (2015), which will, however, not be considered in this thesis.

3.2.3. Covariance and correlation functions

One important characteristic describing the dependence structure of a real-valued random field is its covariance function (or its correlation function, which allows for a more meaningful interpretation). Closely related to the covariance function are the expectation function and the variance function that represent random field analogues to the expectation and variance of a random variable.

Definition 3.2.3 (Expectation function and variance function)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a real-valued random field with domain $\mathcal{T}$ such that $\mathbb{E}\, X^2(t) < \infty$ for all $t \in \mathcal{T}$.

1. The function $\mu_X : \mathcal{T} \to \mathbb{R}$ defined as
$$\mu_X(t) = \mathbb{E}\, X(t), \quad t \in \mathcal{T},$$
is called *expectation function* of X.
2. The function $V_X : \mathcal{T} \to [0, \infty)$ defined as
$$V_X(t) = \operatorname{var} X(t) = \mathbb{E}\left((X(t) - \mu_X(t))^2\right), \quad t \in \mathcal{T},$$
is called *variance function* of X.

In the following, we always assume that if the variance function $V_X : \mathcal{T} \to [0, \infty)$ of a random field $X = \{X(t), t \in \mathcal{T}\}$ exists, then it satisfies that $V_X(t) > 0$ for all $t \in \mathcal{T}$.

Definition 3.2.4 (Covariance function and correlation function)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a real-valued random field with domain $\mathcal{T}$ such that $\mathbb{E}\, X^2(t) < \infty$ for all $t \in \mathcal{T}$.

1. The function $C^\star_X : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ defined as
$$C^\star_X(s,t) = \operatorname{cov}\big(X(s), X(t)\big) = \mathbb{E}\big((X(s) - \mu_X(s))(X(t) - \mu_X(t))\big), \quad s, t \in \mathcal{T},$$
is called *covariance function* of X.
2. The function $\rho^\star_X : \mathcal{T} \times \mathcal{T} \to [-1, 1]$ defined as
$$\rho^\star_X(s,t) = \frac{C^\star_X(s,t)}{\sqrt{V_X(s) \cdot V_X(t)}}, \quad s, t \in \mathcal{T},$$
is called *correlation function* of X.

Remark 3.2.1. Let $X = \{X(t), t \in \mathcal{T}\}$ be a real-valued random field with domain $\mathcal{T}$ such that $\mathbb{E}\, X^2(t) < \infty$ for all $t \in \mathcal{T}$. The covariance function $C_X^\star : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ and the correlation function $\rho_X^\star : \mathcal{T} \times \mathcal{T} \to [-1, 1]$ of X fulfill the following properties:

1. $C_X^\star(t,t) = V_X(t)$ and $\rho_X^\star(t,t) = 1$ for all $t \in \mathcal{T}$,
2. $C_X^\star$ and $\rho_X^\star$ are symmetric, i.e., $C_X^\star(s,t) = C_X^\star(t,s)$ and $\rho_X^\star(s,t) = \rho_X^\star(t,s)$ for all $s, t \in \mathcal{T}$,
3. $C_X^\star$ and $\rho_X^\star$ are positive semi-definite, i.e., for any $n \in \mathbb{N}$, $t_1, \ldots, t_n \in \mathcal{T}$ and $c_1, \ldots, c_n \in \mathbb{R}$ it holds that
$$\sum_{i=1}^{n} \sum_{j=1}^{n} C_X^\star(t_i, t_j) c_i c_j \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_X^\star(t_i, t_j) c_i c_j \geq 0,$$
4. $|C_X^\star(s,t)| \leq \sqrt{V_X(s) \cdot V_X(t)}$, which ensures that $|\rho_X^\star(s,t)| \leq 1$ for all $s, t \in \mathcal{T}$ (Cauchy-Schwarz inequality).

In particular, parametric models for covariance functions, which are frequently used in geostatistical applications such as spatial prediction, need to satisfy these properties in order to be valid. For a proof see Yaglom (1987a), Section 2.

In general, to infer characteristics such as the expectation function and the covariance function of a random field, several realizations of the field need to be available. In most geostatistical datasets, however, only one set of regionalized values is given. In this case, statistical inference is only possible if the underlying random field has a certain homogeneity structure.

Definition 3.2.5 (Strict stationarity and isotropy)**.** A real-valued random field $X = \{X(t), t \in \mathcal{T}\}$ with domain $\mathcal{T}$ is said to be

1. *strictly stationary* if for all $n \in \mathbb{N}$, $t_1, \ldots, t_n \in \mathcal{T}$ and any translation vector $\tau \in \mathbb{R}^d$ such that $t_1 + \tau, \ldots, t_n + \tau \in \mathcal{T}$ it holds that
$$(X(t_1), \ldots, X(t_n)) \stackrel{\mathrm{d}}{=} (X(t_1 + \tau), \ldots, X(t_n + \tau)),$$
2. *strictly isotropic* if for all $n \in \mathbb{N}$, $t_1, \ldots, t_n \in \mathcal{T}$ and any rotation $\delta : \mathbb{R}^d \to \mathbb{R}^d$ around the origin such that $\delta(t_1), \ldots, \delta(t_n) \in \mathcal{T}$ it holds that
$$(X(t_1), \ldots, X(t_n)) \stackrel{\mathrm{d}}{=} (X(\delta(t_1)), \ldots, X(\delta(t_n))).$$

Definition 3.2.6 (Second-order stationarity, isotropy, and motion-invariance)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a real-valued random field with domain $\mathcal{T}$ such that $\mathbb{E}\, X^2(t) < \infty$ for all $t \in \mathcal{T}$. Furthermore, let $\mu_X : \mathcal{T} \to \mathbb{R}$ be the expectation function and $C_X^\star : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ the covariance function of X. Then, X is said to be

1. *second-order stationary* (or weakly stationary) if
   (i) there is some $\mu \in \mathbb{R}$ such that
   $$\mu_X(t) = \mu, \quad t \in \mathcal{T},$$
   (ii) for any translation vector $\tau \in \mathbb{R}^d$ it holds that
   $$C_X^\star(s,t) = C_X^\star(s+\tau, t+\tau), \quad s,t \in \mathcal{T} \text{ such that } s+\tau, t+\tau \in \mathcal{T},$$
2. *second-order isotropic* (or weakly isotropic) if
   (i) for any rotation $\delta : \mathbb{R}^d \to \mathbb{R}^d$ around the origin it holds that
   $$\mu_X(t) = \mu_X(\delta(t)), \quad t \in \mathcal{T} \text{ such that } \delta(t) \in \mathcal{T},$$
   (ii) for any rotation $\delta : \mathbb{R}^d \to \mathbb{R}^d$ around the origin it holds that
   $$C_X^\star(s,t) = C_X^\star(\delta(s), \delta(t)), \quad s,t \in \mathcal{T} \text{ such that } \delta(s), \delta(t) \in \mathcal{T},$$
3. *second-order motion-invariant* (or weakly motion-invariant) if
   (i) X is second-order stationary and isotropic,
   (ii) $C_X^\star$ only depends on the Euclidean distance of its two arguments, i.e.,
   $$C_X^\star(s_1,t_1) = C_X^\star(s_2,t_2), \quad s_1,t_1,s_2,t_2 \in \mathcal{T} \text{ such that } \|s_1 - t_1\|_d = \|s_2 - t_2\|_d.$$

Remark 3.2.2. 1. Provided that $\mathbb{E}\, X^2(t) < \infty$ for all $t \in \mathcal{T}$, one can easily see that a strictly stationary random field $X = \{X(t), t \in \mathcal{T}\}$ is second-order stationary, too. The converse, however, is generally not true. The same relationship also holds for strict and second-order isotropy.

2. If a random field has the domain $\mathcal{T} = \mathbb{R}^d$, then it is second-order stationary and isotropic if and only if it is second-order motion-invariant. However, for $\mathcal{T} \subsetneq \mathbb{R}^d$ a second-order stationary and isotropic random field does not need to be second-order motion-invariant in the sense of Definition 3.2.6 in general.

According to Definition 3.2.6, the following simplified versions of the covariance function and correlation function of a second-order motion-invariant random field can be introduced, which play important roles in geostatistical applications.

Definition 3.2.7 (Motion-invariant covariance and correlation function)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a second-order motion-invariant random field with domain $\mathcal{T}$, covariance function $C_X^\star : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, and correlation function $\rho_X^\star : \mathcal{T} \times \mathcal{T} \to [-1,1]$.

1. The function $C_X : [0, \infty) \to \mathbb{R}$ with

$$C_X(r) = C_X^\star(s,t), \quad s,t \in \mathcal{T} \text{ such that } \|s-t\|_d = r, \tag{3.2}$$

is called the *motion-invariant covariance function* or the *covariogram* of X.

2. The function $\rho_X : [0, \infty) \to [-1, 1]$ with

$$\rho_X(r) = \rho_X^\star(s,t), \quad s,t \in \mathcal{T} \text{ such that } \|s-t\|_d = r, \tag{3.3}$$

is called the *motion-invariant correlation function* or the *correlogram* of X.

If the domain $\mathcal{T}$ is compact, then both C_X and ρ_X are only defined for distances $r \leq r_0$, where r_0 is given according to (3.1).

Remark 3.2.3. The (motion-invariant) covariance function $C_X : [0, \infty) \to \mathbb{R}$ and correlation function $\rho_X : [0, \infty) \to [-1, 1]$ of a second-order motion-invariant random field $X = \{X(t), t \in \mathcal{T}\}$ have the following properties, which follow immediately from Definition 3.2.4, Definition 3.2.7 and Remark 3.2.1:

1. $C_X(0) = V_X(t)$ for all $t \in \mathcal{T}$ and $\rho_X(0) = 1$,
2. $\rho_X(r) = C_X(r)/C_X(0)$ for all $r \geq 0$,
3. for any $n \in \mathbb{N}$, $t_1, \ldots, t_n \in \mathcal{T}$ and $c_1, \ldots, c_n \in \mathbb{R}$ it holds that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} C_X(\|t_i - t_j\|_d) c_i c_j \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_X(\|t_i - t_j\|_d) c_i c_j \geq 0,$$

4. $|C_X(r)| \leq C_X(0)$, which ensures that $|\rho_X(r)| \leq 1$ for all $r \geq 0$ (Cauchy-Schwarz inequality).

Remark 3.2.4 (Great-circle distance)**.** In many geostatistical applications, the domain $\mathcal{T}$ is a subset of the earth's surface and each location $t \in \mathcal{T}$ is identified as $t = (\lambda, \phi)$ with λ and ϕ being the longitude and latitude of t in degree (we call $\mathcal{T}$ a *geographical domain* in this case). Then, formally any location $t \in \mathcal{T}$ is a two-dimensional vector but computing the Euclidean distance $\|s-t\|_2$ between two locations $s, t \in \mathcal{T}$ is not appropriate for two reasons. On the one hand, degrees of longitude and latitude only correspond to approximately equal distances near the equator but distances between two longitudes reduce considerably for increasing or decreasing latitudes. On the other hand, the Euclidean distance of two locations is related to the shortest direct line between those points and is thus expected to underestimate the true distance, which should refer to the shortest path along the earth's surface. This might be negligible if $\mathcal{T}$ represents a small region but, as demonstrated in Banerjee (2005), can influence the results of inference significantly if the domain $\mathcal{T}$ is geographically extensive (e.g., if $\mathcal{T}$

describes the North American continent). In such a case, it is advised to consider the *great-circle distance* $d_{GC}(s,t)$ in km between two locations $s = (\lambda_s, \phi_s), t = (\lambda_t, \phi_t) \in \mathcal{T}$, which is, under the assumption that the earth is a sphere with a radius r_E in km, defined by

$$d_{GC}(s,t) = r_E \cos^{-1}\big(\sin(\phi_s)\sin(\phi_t) + \cos(\phi_s)\cos(\phi_t)\cos(\lambda_s - \lambda_t)\big), \tag{3.4}$$

see, e.g., Diggle and Ribeiro Jr. (2007), Section 2.7. Accordingly, the (motion-invariant) covariance function $C_X : [0,\infty) \to \mathbb{R}$ and correlation function $\rho_X : [0,\infty) \to [-1,1]$ of a second-order motion-invariant random field $X = \{X(t), t \in \mathcal{T}\}$ with geographical domain $\mathcal{T}$ given in (3.2) and (3.3) can also be defined using the great-circle distance $d_{GC}(s,t)$ instead of the two-dimensional Euclidean distance $\|s-t\|_2$.

3.2.4. Cross-covariance and cross-correlation functions

In some geostatistical applications regionalized values of several variables are considered, which are then modeled using more than one random field. In such a case it can be of great interest to not only investigate the correlation structures of the single random fields but also dependencies between them. This can be done, e.g., by means of cross-covariances and cross-correlations. For a more detailed view on this topic, which is generally referred to as multivariate geostatistics, we recommend Wackernagel (2003), Chilès and Delfiner (2012) or Genton and Kleiber (2015).

Definition 3.2.8 (Cross-covariance function and cross-correlation function)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ and $Y = \{Y(t), t \in \mathcal{T}\}$ be real-valued random fields with domain $\mathcal{T}$ such that $\mathbb{E}\, X^2(t), \mathbb{E}\, Y^2(t) < \infty$ for all $t \in \mathcal{T}$.

1. The function $C^\star_{XY} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ defined as

   $$C^\star_{XY}(s,t) = \operatorname{cov}\big(X(s), Y(t)\big) = \mathbb{E}\big((X(s) - \mu_X(s))(Y(t) - \mu_Y(t))\big), \quad s,t \in \mathcal{T},$$

   is called *cross-covariance function* of X and Y.

2. The function $\rho^\star_{XY} : \mathcal{T} \times \mathcal{T} \to [-1,1]$ defined as

   $$\rho^\star_{XY}(s,t) = \frac{C^\star_{XY}(s,t)}{\sqrt{V_X(s) \cdot V_Y(t)}}, \quad s,t \in \mathcal{T},$$

   is called *cross-correlation function* of X and Y.

Definition 3.2.9 (Joint second-order stationarity, isotropy, and motion-invariance)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ and $Y = \{Y(t), t \in \mathcal{T}\}$ be second-order motion-invariant random fields with domain $\mathcal{T}$ and cross-covariance function $C^\star_{XY} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$. Then, X and Y are said to be

1. *jointly second-order stationary* if for any translation vector $\tau \in \mathbb{R}^d$ it holds that

$$C^\star_{XY}(s,t) = C^\star_{XY}(s+\tau, t+\tau), \quad s,t \in \mathcal{T} \text{ such that } s+\tau, t+\tau \in \mathcal{T},$$

2. *jointly second-order isotropic* if for any rotation $\delta : \mathbb{R}^d \to \mathbb{R}^d$ around the origin it holds that

$$C^\star_{XY}(s,t) = C^\star_{XY}(\delta(s), \delta(t)), \quad s,t \in \mathcal{T} \text{ such that } \delta(s), \delta(t) \in \mathcal{T},$$

3. *jointly second-order motion-invariance* if $C^\star_{XY}$ only depends on the Euclidean distance of its two arguments, i.e.,

$$C^\star_{XY}(s_1,t_1) = C^\star_{XY}(s_2,t_2), \quad s_1,t_1,s_2,t_2 \in \mathcal{T} \text{ such that } \|s_1 - t_1\|_d = \|s_2 - t_2\|_d.$$

Remark 3.2.5. 1. Let $X = \{X(t), t \in \mathcal{T}\}$ and $Y = \{Y(t), t \in \mathcal{T}\}$ be real-valued random fields with domain $\mathcal{T}$ such that $\mathbb{E}\, X^2(t), \mathbb{E}\, Y^2(t) < \infty$ for all $t \in \mathcal{T}$. The cross-covariance function $C^\star_{XY} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ and the cross-correlation function $\rho^\star_{XY} : \mathcal{T} \times \mathcal{T} \to [-1,1]$ of X and Y fulfill the following properties:

(i) $C^\star_{XX}(s,t) = C^\star_X(s,t)$ and $\rho^\star_{XX}(s,t) = \rho^\star_X(s,t)$ for all $s,t \in \mathcal{T}$,

(ii) $C^\star_{XY}(s,t) = C^\star_{YX}(t,s)$ and $\rho^\star_{XY}(s,t) = \rho^\star_{YX}(t,s)$ for all $s,t \in \mathcal{T}$ but $C^\star_{XY}$ and $\rho^\star_{XY}$ are generally not symmetric,

(iii) $|C^\star_{XY}(s,t)| \leq \sqrt{V_X(s) \cdot V_Y(t)}$, which ensures that $|\rho^\star_{XY}(s,t)| \leq 1$ for all $s,t \in \mathcal{T}$ (Cauchy-Schwarz inequality).

A proof is similar to the properties stated in Remark 3.2.1.

2. Again, two random fields $X = \{X(t), t \in \mathcal{T}\}$ and $Y = \{Y(t), t \in \mathcal{T}\}$ with domain $\mathcal{T} = \mathbb{R}^d$ are jointly second-order stationary and isotropic if and only if they are jointly second-order motion-invariant. For $\mathcal{T} \subsetneq \mathbb{R}^d$ this is generally not true. In that case, joint second-order stationarity and isotropy could also be defined if both X and Y are second-order stationary and isotropic but not second-order motion-invariant. However, such a scenario is not considered in this thesis.

Definition 3.2.10 (Motion-invariant cross-covariance and cross-correlation function)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ and $Y = \{Y(t), t \in \mathcal{T}\}$ be jointly second-order motion-invariant random fields with domain $\mathcal{T}$, cross-covariance function $C^\star_{XY} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, and cross-correlation function $\rho^\star_{XY} : \mathcal{T} \times \mathcal{T} \to [-1,1]$.

1. The function $C_{XY} : [0,\infty) \to \mathbb{R}$ with

$$C_{XY}(r) = C^\star_{XY}(s,t), \quad s,t \in \mathcal{T} \text{ such that } \|s-t\|_d = r, \tag{3.5}$$

is called the *motion-invariant cross-covariance function* of X and Y.

2. The function $\rho_{XY} : [0, \infty) \to [-1, 1]$ with

$$\rho_{XY}(r) = \rho^{\star}_{XY}(s,t), \quad s,t \in \mathcal{T} \text{ such that } \|s-t\|_d = r, \tag{3.6}$$

is called the *motion-invariant cross-correlation function* of X and Y.

If the domain $\mathcal{T}$ is compact, then both C_{XY} and ρ_{XY} are only defined for distances $r \leq r_0$, where r_0 is given according to (3.1).

Remark 3.2.6. Let $X = \{X(t), t \in \mathcal{T}\}$ and $Y = \{Y(t), t \in \mathcal{T}\}$ be jointly second-order motion-invariant random fields with domain $\mathcal{T}$.

1. The (motion-invariant) cross-covariance function $C_{XY} : [0, \infty) \to \mathbb{R}$ and cross-correlation function $\rho_{XY} : [0, \infty) \to [-1, 1]$ of X and Y have the following properties, which are simple consequences of Definitions 3.2.7, 3.2.8, and 3.2.10 as well as Remarks 3.2.3 and 3.2.5:
   (i) $\rho_{XY}(r) = C_{XY}(r)/\sqrt{C_X(0) \cdot C_Y(0)}$ for all $r \geq 0$,
   (ii) $C_{XY}(r) = C_{YX}(r)$ and $\rho_{XY}(r) = \rho_{YX}(r)$ for all $r \geq 0$,
   (iii) $|C_{XY}(r)| \leq \sqrt{C_X(0) \cdot C_Y(0)}$, which ensures that $|\rho_{XY}(r)| \leq 1$ for all $r \geq 0$ (Cauchy-Schwarz inequality).
2. If $\mathcal{T}$ is a geographical domain, then C_{XY} and ρ_{XY} given in (3.5) and (3.6) can also be defined using the great-circle distance $d_{GC}(s,t)$ introduced in (3.4) instead of the two-dimensional Euclidean distance $\|s-t\|_2$ for $s,t \in \mathcal{T}$.

3.2.5. Semivariograms

An alternative characteristic describing the spatial dependence structure of a random field is the so-called semivariogram. It plays an important role in the context of geostatistical kriging, which is one of the most commonly used spatial interpolation methods. We put a particular emphasis on a motion-invariant version of the semivariogram, where it can be shown that, basically, the semivariogram contains the same information as the covariance function.

Definition 3.2.11 (Semivariogram)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a real-valued random field with domain $\mathcal{T}$ such that $\mathbb{E}\, X^2(t) < \infty$ for all $t \in \mathcal{T}$. We call the function $\gamma^{\star}_X : \mathcal{T} \times \mathcal{T} \to [0, \infty)$ defined by

$$\gamma^{\star}_X(s,t) = \frac{1}{2} \operatorname{var}\big(X(s) - X(t)\big), \quad s,t \in \mathcal{T},$$

the *semivariogram* of X.

Remark 3.2.7. One can show that the semivariogram $\gamma_X^\star : \mathcal{T}\times\mathcal{T} \to [0,\infty)$ of a random field $X = \{X(t), t \in \mathcal{T}\}$ is closely related to the covariance function $C_X^\star : \mathcal{T}\times\mathcal{T} \to \mathbb{R}$ of X. In particular, we get that

$$\gamma_X^\star(s,t) = \frac{1}{2}\operatorname{var}\big(X(s) - X(t)\big) = \frac{1}{2}\big(V_X(s) + V_X(t)\big) - C_X^\star(s,t), \quad s,t \in \mathcal{T}.$$

If X is second-order motion-invariant, this further simplifies to

$$\gamma_X^\star(s,t) = C_X(0) - C_X(\|s-t\|_d), \quad s,t \in \mathcal{T},$$

which implies that $\gamma_X^\star(s,t)$ only depends on the d-dimensional Euclidean distance $\|s-t\|_d$ of $s,t \in \mathcal{T}$. This motivates the definition of a motion-invariant version of the semivariogram $\gamma_X^\star$, which is one of the most important tools used in geostatistics.

Definition 3.2.12 (Motion-invariant semivariogram)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a second-order motion-invariant random field with domain $\mathcal{T}$ and semivariogram $\gamma_X^\star : \mathcal{T}\times\mathcal{T} \to [0,\infty)$. The function $\gamma_X : [0,\infty) \to [0,\infty)$ given by

$$\gamma_X(r) = \gamma_X^\star(s,t), \quad s,t \in \mathcal{T} \text{ such that } \|s-t\|_d = r, \tag{3.7}$$

is called the *motion-invariant semivariogram* of X. If the domain $\mathcal{T}$ is compact, then γ_X is only defined for distances $r \le r_0$, where r_0 is given according to (3.1).

Remark 3.2.8. Let $X = \{X(t), t \in \mathcal{T}\}$ be a second-order motion-invariant random field with domain $\mathcal{T}$ and (motion-invariant) semivariogram $\gamma_X : [0,\infty) \to [0,\infty)$.

1. The semivariogram γ_X has the following properties:
   (i) $\gamma_X(0) = 0$,
   (ii) $\gamma_X(r) = C_X(0) - C_X(r)$ for all $r \ge 0$,
   (iii) γ_X is conditionally negative semi-definite, i.e., for any $n \in \mathbb{N}$, $t_1, \ldots, t_n \in \mathcal{T}$ and $c_1, \ldots, c_n \in \mathbb{R}$ with $\sum_{i=1}^n c_i = 0$ it holds that

   $$\sum_{i=1}^n \sum_{j=1}^n \gamma_X(\|t_i - t_j\|_d) c_i c_j \le 0.$$

   The first and second property follow immediately from Definition 3.2.11, Definition 3.2.12, and Remark 3.2.7. For a proof of the third property see, e.g., Montero et al. (2015), Section 3.4.1.

2. If $\mathcal{T}$ is a geographical domain, then γ_X given in (3.7) can also be defined using the great-circle distance $d_{GC}(s,t)$ introduced in (3.4) instead of the two-dimensional Euclidean distance $\|s-t\|_2$ for $s,t \in \mathcal{T}$.

3. The behavior of the semivariogram γ_X is often characterized by three parameters. As mentioned above, it holds that $\gamma_X(0) = 0$ but in many applications it is required (e.g., due to being observed in data) that $\gamma_X(h)$ tends to some positive constant $c_0 > 0$ as $h \to 0$. This implies that γ_X has a discontinuity in $r = 0$, which is called *nugget effect*. The parameter $c_0 > 0$ is referred to as *nugget*. Furthermore, it is often assumed that $C_X(r) \to 0$ for $r \to \infty$ as, typically, values $X(s)$ and $X(t)$ of the random field X at $s, t \in \mathcal{T}$ are uncorrelated if the distance $\|s - t\|_d$ is large enough. This implies that $\gamma_X(r) \to C_X(0)$ as $r \to \infty$, i.e., $C_X(0)$ is the limiting value of γ_X is denoted as *sill*. The difference between the sill $C_X(0)$ and the nugget c_0 is called *partial sill*. The third parameter of interest, the *range* $a > 0$ of γ_X, denotes the smallest distance $r \geq 0$ such that $\gamma_X(r')$ is equal to the sill $C_X(0)$ for each $r' \geq r$. If $\gamma_X(r) < C_X(0)$ for all $r \geq 0$ then the range is often defined as the smallest distance $r \geq 0$ such that $\gamma_X(r')$ is greater than $u \cdot C_X(0)$ for each $r' \geq r$, where the threshold $u \in (0, 1)$ is set to, e.g., $u = 0.95$ or $u = 0.99$.

Example 3.2.2 (Semivariogram models)**.** Since commonly used estimators for the (motion-invariant) semivariogram $\gamma_X : [0, \infty) \to [0, \infty)$ of a second-order motion-invariant random field $X = \{X(t), t \in \mathcal{T}\}$ do typically not satisfy the theoretical properties of γ_X (and have other disadvantages, see Section 3.2.6), parametric semivariogram models are often fitted in applications. Three popular parametric models are introduced in the following, for further examples see Cressie (1993), Jian et al. (1996) or Montero et al. (2015).

1. If

$$\gamma_X(r) = \begin{cases} 0, & r = 0, \\ c_0 + c\left(\frac{3}{2}\frac{r}{a} - \frac{1}{2}\left(\frac{r}{a}\right)^3\right), & 0 < r < a, \\ c_0 + c, & r \geq a, \end{cases}$$

then γ_X is called a *spherical semivariogram* with parameters $c_0 \geq 0$ and $c, a > 0$. The nugget of γ_X is given by c_0, the sill by $c_0 + c$, and the range by a. Note that the spherical model is a valid semivariogram for dimension $d \leq 3$ only.

2. If

$$\gamma_X(r) = \begin{cases} 0, & r = 0, \\ c_0 + c\left(7\left(\frac{r}{a}\right)^2 - \frac{35}{4}\left(\frac{r}{a}\right)^3 + \frac{7}{2}\left(\frac{r}{a}\right)^5 - \frac{3}{4}\left(\frac{r}{a}\right)^7\right), & 0 < r < a, \\ c_0 + c, & r \geq a, \end{cases}$$

then γ_X is called a *cubic semivariogram* with parameters $c_0 \geq 0$ and $c, a > 0$. The nugget of γ_X is given by c_0, the sill by $c_0 + c$ and the range by a. Note that the cubic model, like the spherical model, is a valid semivariogram for dimension $d \leq 3$ only.

3. If

$$\gamma_X(r) = \begin{cases} 0, & r = 0, \\ c_0 + c\left(1 - \exp\left\{-\frac{r}{a_e}\right\}\right), & r > 0, \end{cases}$$

then γ_X is called an *exponential semivariogram* with parameters $c_0 \geq 0$ and $c, a_e > 0$. The nugget of γ_X is given by c_0 and the sill by $c_0 + c$. The range a, however, is more difficult to determine since $\gamma_X(r) \to c_0 + c$ for $r \to \infty$ but $\gamma_X(r) < c_0 + c$ for each $r \geq 0$. In Montero et al. (2015), Section 3.5.1.2 it is suggested to determine the range a using the threshold $u = 0.95$, i.e., a should satisfy that $\gamma_X(r') \geq 0.95\,(c_0 + c)$ for each $r' \geq a$, which is fulfilled for

$$a = a_e \log\left(20\left(1 + \frac{c_0}{c}\right)^{-1}\right).$$

Remark 3.2.9. Analogous to the cross-covariance function of two random fields $X = \{X(t), t \in \mathcal{T}\}$ and $Y = \{Y(t), t \in \mathcal{T}\}$, a cross-semivariogram of X and Y can be defined. Besides the general version, which does not require any conditions on the homogeneity of X and Y but is difficult to handle in practice, a motion-invariant version can be given if X and Y are jointly second-order motion-invariant. However, cross-semivariograms are not considered in this thesis, for more details see Wackernagel (2003), Chapter 20 and Chilès and Delfiner (2012), Section 5.6.2.

3.2.6. Estimators for (cross-) correlation functions and semivariograms

Now that several characteristics have been introduced which describe dependencies in (or between) random fields, we address the question how these characteristics can be estimated based on observed data. In applications, typically regionalized values $x(t_1), \ldots, x(t_n)$ are given for a finite set of $n \in \mathbb{N}$ sampling points $t_1, \ldots, t_n$ in a domain $\mathcal{T}$ as specified in Section 3.2.2. These data are interpreted as realizations of a random field $X = \{X(t), t \in \mathcal{T}\}$ at $t_1, \ldots, t_n$ and it is usually assumed that X is second-order motion-invariant as otherwise, estimation based on only one set of regionalized values is considerably more complicated (or even impossible). If the domain $\mathcal{T}$ is compact, then the estimators of (cross-) covariance functions, (cross-) correlation functions and semivariograms introduced in this section are only defined for distances $r \leq r_0$, where r_0 is given according to (3.1). We start with the discussion of estimators for (cross-) covariance and (cross-) correlation functions. The most intuitive approach relies on the method of moments.

Definition 3.2.13 (Method of moment estimator for (cross-) covariance function)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a second-order motion-invariant random field with domain $\mathcal{T}$, constant expectation $\mu_X \in \mathbb{R}$, and (motion-invariant) covariance function C_X :

$[0, \infty) \to \mathbb{R}$. Furthermore, let $t_1, \ldots, t_n \in \mathcal{T}$ with $n \in \mathbb{N}$ be a fixed sequence of sampling points and define

$$N(r) = \{(t_i, t_j) : \|t_i - t_j\|_d = r; i, j = 1, \ldots, n\}, \quad r \geq 0. \tag{3.8}$$

1. The estimator $\hat{\mu}_X$ defined by

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^{n} X(t_i) \tag{3.9}$$

is a *method of moment estimator* of the expectation μ_X, which is also denoted as *empirical mean*.

2. The estimator $\hat{C}_X^M$ defined by

$$\hat{C}_X^M(r) = \frac{1}{\#N(r)} \sum_{(t_i,t_j)\in N(r)} (X(t_i) - \hat{\mu}_X)(X(t_j) - \hat{\mu}_X), \quad r \geq 0 \text{ such that } \#N(r) > 0, \tag{3.10}$$

is a *method of moment estimator* of the covariance function C_X, which is also denoted as *empirical* or *experimental covariogram*.

3. Let $Y = \{Y(t), t \in \mathcal{T}\}$ be another second-order motion-invariant random field with domain $\mathcal{T}$ and constant expectation $\mu_Y \in \mathbb{R}$ such that X and Y are jointly second-order motion-invariant with (motion-invariant) cross-covariance function $C_{XY} : [0, \infty) \to \mathbb{R}$. The estimator $\hat{C}_{XY}^M$ defined by

$$\hat{C}_{XY}^M(r) = \frac{1}{\#N(r)} \sum_{(t_i,t_j)\in N(r)} (X(t_i) - \hat{\mu}_X)(Y(t_j) - \hat{\mu}_Y), \quad r \geq 0 \text{ such that } \#N(r) > 0, \tag{3.11}$$

is a *method of moment estimator* of the cross-covariance function C_{XY}.

Remark 3.2.10. Let $X = \{X(t), t \in \mathcal{T}\}$ and $Y = \{Y(t), t \in \mathcal{T}\}$ be given as in Definition 3.2.13.

1. In most applications there are no or only few distances $r \geq 0$ such that $\#N(r) > 1$, which results in unreliable estimates of (cross-) covariance functions. Thus, it is common to also add those pairs (t_i, t_j) to $N(r)$ that have a distance $\|t_i - t_j\|_d$ of approximately r. For that purpose, we consider a fixed tolerance value $\varepsilon > 0$ and define

$$N^\varepsilon(r) = \{(t_i, t_j) : \|t_i - t_j\|_d \in [r - \varepsilon, r + \varepsilon]; i, j = 1, \ldots, n\}, \quad r \geq 0, \tag{3.12}$$

which is often used instead of $N(r)$ in (3.10) and (3.11) (this is sometimes denoted as *binning* in literature). The interval $[r - \varepsilon, r + \varepsilon]$ is called a *tolerance interval* in this context. A usual proceeding is to estimate a (cross-) covariance function using binning at a finite sequence $r_1, \dots, r_m \geq 0$ of distances with $m \in \mathbb{N}$, where the tolerance value ε and the distances $r_1, \dots, r_m$ are chosen in such a way that the tolerance intervals $[r_1 - \varepsilon, r_1 + \varepsilon]$, ..., $[r_m - \varepsilon, r_m + \varepsilon]$ form a partition of the interval $[r_1 - \varepsilon, r_m + \varepsilon]$.

2. The empirical mean $\hat{\mu}_X$ as introduced in (3.9) is an unbiased estimator of the constant expectation μ_X, i.e., $\mathbb{E}\,\hat{\mu}_X = \mu_X$ but due to $X(t_1), \dots, X(t_n)$ being not independent in general, it is not the estimator with minimum variance. In order to construct an estimator with smaller variance, the covariance function $C_X : [0, \infty) \to \mathbb{R}$ is needed to be known which is typically not the case in applications, see Montero et al. (2015), Section 3.3.

3. The method of moments estimator $\hat{C}_X^M$ given in (3.10) is biased, i.e., in general $\mathbb{E}\,\hat{C}_X^M(r) \neq C_X(r)$ for $r \geq 0$. The bias can have a substantial influence if n is small (Cressie, 1993, Section 2.4.1). Furthermore, estimates obtained according to $\hat{C}_X^M$ tend to be unstable (i.e., large jumps often occur) and are generally not positive semi-definite, which implies that they should not be used for spatial prediction (Montero et al., 2015, Section 3.3). The same applies for the method of moments estimator $\hat{C}_{XY}^M$ of the cross-covariance function C_{XY} given in (3.11).

4. If $\mathcal{T}$ is a geographical domain, then $N(r)$ and $N^{\varepsilon}(r)$ given in (3.8) and (3.12) can also be defined using the great-circle distance $d_{GC}(t_i, t_j)$ introduced in (3.4) instead of the two-dimensional Euclidean distance $\|t_i - t_j\|_2$ for $t_i, t_j \in \mathcal{T}$.

As stated in Remark 3.2.10, the proposed method of moments estimator for (cross-) covariance functions has certain disadvantages, which do not only impede a direct use in geostatistical applications but also make analysis and interpretation of obtained estimates difficult. A popular solution to these shortcomings is to fit a parametric model to estimates obtained by the method of moments, see, e.g., Montero et al. (2015), Section 3.2.2. An alternative approach is to use nonparametric estimators as discussed in, e.g., Hall et al. (1994), Genton and Gorsich (2002) or Choi et al. (2013). In this thesis, we consider a kernel-based estimator suggested in Hall et al. (1994).

Definition 3.2.14 (Kernel estimator for (cross-) covariance function)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a second-order motion-invariant random field with domain $\mathcal{T}$ and (motion-invariant) covariance function $C_X : [0, \infty) \to \mathbb{R}$. Furthermore, let $t_1, \dots, t_n \in \mathcal{T}$ with $n \in \mathbb{N}$ be a fixed sequence of sampling points, $h > 0$ a positive constant and $\kappa : \mathbb{R} \to [0, \infty)$ a nonnegative, bounded, and Borel-measurable function with $\int_{\mathbb{R}} \kappa(x)\,\mathrm{d}x = 1$, $\int_{\mathbb{R}} x\kappa(x)\,\mathrm{d}x = 0$, and $\int_{\mathbb{R}} x^2\kappa(x)\,\mathrm{d}x < \infty$ (called a kernel function, see Section 3.4.1).

1. The estimator $\hat{C}_X^K$ defined by

$$\hat{C}_X^K(r) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(X(t_i) - \hat{\mu}_X)(X(t_j) - \hat{\mu}_X)\,\kappa\left(\frac{r-\|t_i-t_j\|_d}{h}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa\left(\frac{r-\|t_i-t_j\|_d}{h}\right)}, \quad r \geq 0, \tag{3.13}$$

   is a *kernel estimator* of the covariance function C_X with bandwidth $h > 0$.

2. Let $Y = \{Y(t), t \in \mathcal{T}\}$ be another second-order motion-invariant random field with domain $\mathcal{T}$ such that X and Y are jointly second-order motion-invariant with (motion-invariant) cross-covariance function $C_{XY} : [0, \infty) \to \mathbb{R}$. The estimator $\hat{C}_{XY}^K$ defined by

$$\hat{C}_{XY}^K(r) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(X(t_i) - \hat{\mu}_X)(Y(t_j) - \hat{\mu}_Y)\,\kappa\left(\frac{r-\|t_i-t_j\|_d}{h}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa\left(\frac{r-\|t_i-t_j\|_d}{h}\right)}, \quad r \geq 0, \tag{3.14}$$

   is a *kernel estimator* of the cross-covariance function C_{XY} with bandwidth $h > 0$.

If for any $r \geq 0$ the denominator in (3.13) is equal to zero, then the corresponding numerator is equal to zero as well. In this case, the estimator $\hat{C}_X^K(r)$ is not defined. The same applies for (3.14).

Remark 3.2.11. Let $X = \{X(t), t \in \mathcal{T}\}$ and $Y = \{Y(t), t \in \mathcal{T}\}$ be given as in Definition 3.2.14.

1. The bandwidth $h > 0$ in (3.13) and (3.14) controls the degree of smoothness of estimated (cross-) covariance functions. It is a complicated task to choose the bandwidth in such a way that spurious random variations are removed without eliminating important details in the data. This can be done either manually, by comparing estimates for different choices of h, or automatically. Some algorithmic approaches for bandwidth selection are discussed in Section 3.4.4.

2. If $\mathcal{T}$ is a geographical domain, then the kernel estimators $\hat{C}_X^K$ and $\hat{C}_{XY}^K$ given in (3.13) and (3.14) can also be defined using the great-circle distance $d_{GC}(t_i, t_j)$ introduced in (3.4) instead of the two-dimensional Euclidean distance $\|t_i - t_j\|_2$ for $t_i, t_j \in \mathcal{T}$.

Using the relationship between (cross-) covariance and (cross-) correlation functions specified in Remarks 3.2.3 and 3.2.6, the following plug-in estimators for (motion-invariant) correlation and cross-correlation functions can be given.

Definition 3.2.15 (Plug-in estimator for (cross-) correlation function)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a second-order motion-invariant random field with domain $\mathcal{T}$, (motion-invariant) covariance function $C_X : [0,\infty) \to \mathbb{R}$, and (motion-invariant) correlation function $\rho_X : [0,\infty) \to [-1,1]$. Furthermore, let $\hat{C}_X$ be an estimator of C_X, with $\hat{C}_X(0) > 0$.

1. The estimator $\hat{\rho}_X$ defined by

$$\hat{\rho}_X(r) = \frac{\hat{C}_X(r)}{\hat{C}_X(0)}, \quad r \geq 0, \tag{3.15}$$

   is a *plug-in estimator* of the correlation function ρ_X.

2. Let $Y = \{Y(t), t \in \mathcal{T}\}$ be another second-order motion-invariant random field with domain $\mathcal{T}$ and (motion-invariant) covariance function $C_Y : [0,\infty) \to \mathbb{R}$ such that X and Y are jointly second-order motion-invariant with (motion-invariant) cross-covariance function $C_{XY} : [0,\infty) \to \mathbb{R}$ and cross-correlation function $\rho_{XY} : [0,\infty) \to [-1,1]$. Furthermore, let $\hat{C}_Y$ be an estimator of C_Y with $\hat{C}_Y(0) > 0$ and $\hat{C}_{XY}$ an estimator of C_{XY}. The estimator $\hat{\rho}_{XY}$ defined by

$$\hat{\rho}_{XY}(r) = \frac{\hat{C}_{XY}(r)}{\sqrt{\hat{C}_X(0) \cdot \hat{C}_Y(0)}}, \quad r \geq 0, \tag{3.16}$$

   is a *plug-in estimator* of the cross-correlation function ρ_{XY}.

Finally, a method of moment estimator for the motion-invariant semivariogram of a second-order motion-invariant random field is discussed.

Definition 3.2.16 (Method of moment estimator for semivariogram)**.** Let $X = \{X(t), t \in \mathcal{T}\}$ be a second-order motion-invariant random field with domain $\mathcal{T}$ and (motion-invariant) semivariogram $\gamma_X : [0,\infty) \to [0,\infty)$. Furthermore, let $t_1, \ldots, t_n \in \mathcal{T}$ with $n \in \mathbb{N}$ be a fixed sequence of sampling points and for each $r \geq 0$ let $N(r)$ be given as in (3.8). The estimator $\hat{\gamma}_X$ defined by

$$\hat{\gamma}_X(r) = \frac{1}{2\#N(r)} \sum_{(t_i,t_j)\in N(r)} \big(X(t_i) - X(t_j)\big)^2, \quad r \geq 0 \text{ such that } \#N(r) > 0, \tag{3.17}$$

is a *method of moment estimator* of the semivariogram γ_X, which is also denoted as *empirical* or *experimental semivariogram.*

Remark 3.2.12. Let $X = \{X(t), t \in \mathcal{T}\}$ be a second-order motion-invariant random field with domain $\mathcal{T}$, (motion-invariant) semivariogram $\gamma_X : [0,\infty) \to [0,\infty)$, and (motion-invariant) covariance function $C_X : [0,\infty) \to \mathbb{R}$.

1. Since γ_X can also be represented as

$$\gamma_X(r) = \frac{1}{2}\,\mathbb{E}\left(X(s) - X(t)\right)^2, \quad s, t \in \mathcal{T} \text{ such that } \|s - t\|_d = r,$$

   one can easily see that $\mathbb{E}\,\hat{\gamma}_X(r) = \gamma_X(r)$ for $r \geq 0$ (such that $\#N(r) > 0$), i.e., $\hat{\gamma}_X$ is unbiased (in contrast to the method of moment estimator $\hat{C}_X^M$ of C_X). The reason for having this desirable property is that the unknown constant expectation $\mu_X \in \mathbb{R}$ of X is not involved in the definition of the semivariogram. Due to $X(t_1), \ldots, X(t_n)$ being not independent in general, the derivation of further characteristics (such as second moments) of the estimator $\hat{\gamma}_X$ is only possible under restrictive assumptions on X, see Cressie (1993), Section 2.4 and Schabenberger and Gotway (2005), Section 4.4.1.

2. When applying the estimator $\hat{\gamma}_X$ in practice, it is customary to use binning, compare to the discussion of the method of moment estimator $\hat{C}_X^M$ in Remark 3.2.10. This means that $\hat{\gamma}_X$ is often computed for a finite sequence $r_1, \ldots, r_m \geq 0$ of distances with $m \in \mathbb{N}$, where $N(r_j)$ is replaced by the extended set $N^\varepsilon(r_j)$ as introduced in (3.12) for $j = 1, \ldots, m$. Typically, the distances $r_1, \ldots, r_m$ and the tolerance value $\varepsilon > 0$ are chosen in such a way that $[r_1 - \varepsilon, r_1 + \varepsilon], \ldots, [r_m - \varepsilon, r_m + \varepsilon]$ form a partition of the interval $[r_1 - \varepsilon, r_m + \varepsilon]$. Moreover, in Journel and Huijbregts (1978) it is advised that each $N^\varepsilon(r_j)$ should contain at least 30 (better 50) pairs of sample points for $j = 1, \ldots, m$. However, regardless of whether binning is used or not, obtained estimates are not conditionally negative semi-definite in general and tend to be unstable, showing large jumps occasionally.

3. The method of moment estimator $\hat{\gamma}_X$ is not very robust to outliers (Cressie, 1993, Section 2.2.1). Therefore, several more robust approaches to the estimation of γ_X are suggested in the literature, see, e.g., Cressie (1993), Section 2.4.3. However, most of them do not account for the disadvantages of $\hat{\gamma}_X$ mentioned previously.

Similar as for the covariance function, nonparametric estimators are proposed in literature, which provide smooth (but not necessarily conditionally negative semi-definite) estimates of the semivariogram, see, e.g., Schabenberger and Gotway (2005), Section 4.6, Diggle and Ribeiro Jr. (2007), Section 5.2.3 or Kim and Park (2012). In the present thesis, however, such approaches are not discussed.

3.2.7. Fitting of semivariogram models

Variogram estimates obtained using the method of moments have several disadvantages that do not only make analysis and interpretation difficult but also impede an application in spatial prediction. Thus, a popular approach in geostatistics is to fit a parametric semivariogram model that matches the dependence structure in the available data as well as possible. Several fitting methods are considered in the literature with least squares, maximum likelihood, and composite likelihood being the most commonly used. For a general overview of semivariogram fitting we refer to Cressie (1993), Section 2.6, Schabenberger and Gotway (2005), Section 4.5, Chilès and Delfiner (2012), Section 2.6, and Montero et al. (2015), Section 3.8.

Let $X = \{X(t), t \in \mathcal{T}\}$ be a second-order motion-invariant random field with domain $\mathcal{T}$. We suppose that the motion-invariant semivariogram of X belongs to a parametric family of semivariogram models, which is why we use the notation $\gamma_X^\theta : [0,\infty) \to [0,\infty)$ with some parameter vector $\theta \in \Theta \subset \mathbb{R}^k$ and some $k \in \mathbb{N}$. For example, γ_X^θ could describe a spherical, cubic or exponential semivariogram model as introduced in Example 3.2.2, where in all three cases $\Theta = [0,\infty) \times (0,\infty)^2$ and $k = 3$, but many more classes of parametric models exist. Furthermore, let $t_1, \ldots, t_n \in \mathcal{T}$ with $n \in \mathbb{N}$ be a fixed sequence of sampling points. We describe how the parameter vector θ can be fitted based on $X(t_1), \ldots, X(t_n)$ using a generalized least squares (GLS) method with iterative reweighting. Note that ordinary least squares (OLS) should not be applied here since the random variables $X(t_1), \ldots, X(t_n)$ are not independent in general. Besides the references given above, we recommend Cressie (1985) and Müller (1999) for a more detailed discussion of this method.

As a basis for model fitting, we consider a sequence $r_1, \ldots, r_m \geq 0$ of distances with $m \in \mathbb{N}$ and the corresponding method of moment estimators $\hat{\gamma}_X(r_1), \ldots, \hat{\gamma}_X(r_m)$ that are given according to (3.17), where typically binning is applied. In the following, we write $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_X(r_1), \ldots, \hat{\gamma}_X(r_m))^\top$ and $\boldsymbol{\gamma}^{\boldsymbol{\theta}} = (\gamma_X^\theta(r_1), \ldots, \gamma_X^\theta(r_m))^\top$. We suppose that

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^{\boldsymbol{\theta}} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} : \Omega \to \mathbb{R}^m$ is a centered random vector (i.e., all its components have an expectation of zero) with covariance matrix $\Sigma \in \mathbb{R}^{m \times m}$ (which also is the covariance matrix of $\hat{\boldsymbol{\gamma}}$). Then, the GLS estimator $\hat{\theta}_{GLS}$ of θ is given by

$$\hat{\theta}_{GLS} = \underset{\theta \in \Theta}{\arg\min} \left\{ (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^{\boldsymbol{\theta}})^\top \Sigma^{-1} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^{\boldsymbol{\theta}}) \right\}. \tag{3.18}$$

The biggest challenge is the determination of the covariance matrix Σ as no general formula is known in the literature. However, under the assumption that X is a Gaussian random field (i.e., all its finite-dimensional distributions belong to the family of multivariate normal distributions), it holds that

1. $\mathbb{E}\left((X(s) - X(t))^2\right) = 2\gamma_X^\theta(\|s-t\|_d), \quad s, t \in \mathcal{T},$

2. $\operatorname{var}\left((X(s) - X(t))^2\right) = 8\left(\gamma_X^\theta(\|s-t\|_d)\right)^2, \quad s, t \in \mathcal{T},$

3. $\operatorname{cov}\left((X(s_1) - X(t_1))^2, (X(s_2) - X(t_2))^2\right)$

$$= 2\left(\gamma_X^\theta(\|s_1 - t_2\|_d) + \gamma_X^\theta(\|t_1 - s_2\|_d) - \gamma_X^\theta(\|s_1 - s_2\|_d) - \gamma_X^\theta(\|t_1 - t_2\|_d)\right)^2, \qquad s_1, s_2, t_1, t_2 \in \mathcal{T},$$

see Montero et al. (2015), Section 3.6, which can be used to give a (complicated) closed formula for the covariance matrix Σ, see Cressie (1985). Note that in this case, Σ also depends on the parameter vector θ, which is emphasized by writing $\Sigma(\theta)$ in the following. Therefore, it does not seem possible to directly compute $\hat{\theta}_{GLS}$ according to (3.18) in one step. Instead, the following iterative algorithm for the estimation of the parameter vector θ is proposed in Müller (1999).

Algorithm 3.2.1.

1. Set $l = 0$ and $\Sigma = I$ and compute an initial estimator $\hat{\theta}_{GLS}^{(0)}$ according to (3.18). In this case, $\hat{\theta}_{GLS}^{(0)}$ actually is an OLS estimator.
2. Compute the covariance matrix $\Sigma(\hat{\theta}_{GLS}^{(l)})$ using the moment formulas given above based on $\hat{\theta}_{GLS}^{(l)}$ and set $\Sigma = \Sigma(\hat{\theta}_{GLS}^{(l)})$.
3. Compute the estimator $\hat{\theta}_{GLS}^{(l+1)}$ according to (3.18). If the Euclidean norm $\|\hat{\theta}_{GLS}^{(l+1)} - \hat{\theta}_{GLS}^{(l)}\|_2$ is small enough, then set $l = l+1$ and terminate the algorithm. Otherwise, set $l = l + 1$ and go back to step 2.

After the algorithm terminates, $\hat{\theta}_{GLS}^{(l)}$ is used as an estimator of the parameter vector θ.

3.3. Stochastic geometry

In the present section, we discuss some basic models from stochastic geometry including important properties, characteristics, and simulation algorithms. According to Chiu et al. (2013), stochastic geometry is the field of mathematical research that provides models and methods for the analysis of complicated geometrical structures occurring in datasets from different areas in science and technology. In this thesis, we focus on random point processes and random germ-grain models although a wide range of more general models exists. For a more detailed overview of this topic we recommend Stoyan and Stoyan (1994), Beneš and Rataj (2004), Baddeley et al. (2007), Schneider and Weil (2008), and Chiu et al. (2013).

3.3.1. Random measures

Random measures are some of the most important and most flexible tools considered in stochastic geometry (Kallenberg, 1986), which allow to formally describe stochastic models for random geometric objects in a particularly elegant way. Moreover, many models of stochastic geometry have characteristics that are associated with random measures, which further emphasizes their importance. Let M denote the family of all locally finite measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where a measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ is said to be locally finite if $\mu(B) < \infty$ for all $B \in \mathcal{B}_0(\mathbb{R}^d)$. Furthermore, let $\mathcal{M}$ be the smallest σ-algebra of subsets of M such that for each $B \in \mathcal{B}_0(\mathbb{R}^d)$ the mapping that assigns the value $\mu(B)$ to each measure $\mu \in \mathsf{M}$ is $(\mathcal{M}, \mathcal{B}(\mathbb{R}))$-measurable.

Definition 3.3.1 (Random measure)**.** A *random measure* $M : \Omega \to \mathsf{M}$ is a random element defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in the measurable space $(\mathsf{M}, \mathcal{M})$.

Remark 3.3.1. A random measure M can also be interpreted as a random function with index set $\mathcal{T} = \mathcal{B}(\mathbb{R}^d)$, compare to Example 3.2.1. Thus, we often use the notation $M = \{M(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ in this thesis, where $M(B)$ denotes the random value of M for each Borel set $B \in \mathcal{B}(\mathbb{R}^d)$. According to Theorem 3.2.2, the distribution of M is uniquely determined by its family of finite-dimensional distributions.

Definition 3.3.2 (Intensity measure)**.** Let $M = \{M(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ be a random measure defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The (deterministic) *intensity measure* $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ of M is defined as

$$\mu(B) = \mathbb{E}\, M(B), \quad B \in \mathcal{B}(\mathbb{R}^d).$$

Remark 3.3.2. In general, it is not guaranteed that the intensity measure μ of a (locally finite) random measure M is locally finite, although this is assumed in most applications.

Definition 3.3.3 (Stationarity and isotropy)**.** A random measure $M = \{M(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ is said to be

1. *stationary* if for all $n \in \mathbb{N}$, $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R}^d)$ and any translation vector $\tau \in \mathbb{R}^d$ it holds that

$$(M(B_1), \ldots, M(B_n)) \stackrel{\mathrm{d}}{=} (M(B_1 + \tau), \ldots, M(B_n + \tau)),$$

2. *isotropic* if for all $n \in \mathbb{N}$, $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R}^d)$ and any rotation $\delta : \mathbb{R}^d \to \mathbb{R}^d$ around the origin it holds that

$$(M(B_1), \ldots, M(B_n)) \stackrel{\mathrm{d}}{=} (M(\delta B_1), \ldots, M(\delta B_n)).$$

3.3.2. Random point processes

Random point processes are basic tools of stochastic geometry that are frequently used for the modeling of irregular point patterns. If considered from an applied perspective, random point processes are often interpreted as random configurations of points in space and/or time but a more formal description using random counting measures is typically preferred in the scientific literature. Historically, point processes were introduced for the modeling of random sequences of temporal events in dimension $d = 1$. Since for higher dimensions often no temporal evolution is considered, the term random point field would be more appropriate in that case (Chiu et al., 2013). However, the notation point process is still used by most researchers for arbitrary dimensions $d \in \mathbb{N}$. Besides the references given at the beginning of Section 3.3, we recommend König and Schmidt (1992), Daley and Vere-Jones (2003), Daley and Vere-Jones (2008), Illian et al. (2008), and Diggle (2014) for a more detailed view on this topic.

Definition 3.3.4 (Random point process)**.** Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^d \cup \{\infty\}$ be an arbitrary sequence of random vectors defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that

$$\mathbb{P}\big(\#\{i : X_i \in B\} < \infty\big) = 1, \quad B \in \mathcal{B}_0(\mathbb{R}^d),$$

and

$$\mathbb{P}\big(\{X_i \neq X_j\} \cup \{X_i = X_j = \infty\}\big) = 1, \quad i, j \in \mathbb{N}, i \neq j.$$

Then, $X = \{X_i, i = 1, 2, \ldots\}$ is called a (locally finite and simple) *random point process*.

Remark 3.3.3. The notation $X_i = \infty$ means that the i-th point X_i of a random point process $X = \{X_i, i = 1, 2, \ldots\}$ does not exist, which allows for the description of point processes whose realizations have a finite number of points. Alternatively, one can also write $\{X_i, i = 1, \ldots, Z\}$ where $Z : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ is a random variable describing the total number of points of X in $\mathbb{R}^d$, i.e., $Z = \#\{i : X_i \in \mathbb{R}^d\}$. This notation is particularly useful when a random point process is restricted to some compact subset $W \subset \mathbb{R}^d$ or for the construction of so-called cluster processes, see Section 3.3.5.

As indicated previously, random point processes can also be interpreted as random counting measures, which facilitates a formal description of point process models and their properties. A measure $\varphi \in \mathsf{M}$ is called a (locally finite) counting measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ if $\varphi(B) \in \mathbb{N}_0 \cup \{\infty\}$ for all $B \in \mathcal{B}(\mathbb{R}^d)$. Let N denote the family of all locally finite and simple counting measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where a counting measure $\varphi : \mathcal{B}(\mathbb{R}^d) \to \mathbb{N}_0 \cup \{\infty\}$ is said to be simple if $\varphi(\{x\}) \in \{0, 1\}$ for each $x \in \mathbb{R}^d$. Furthermore, let $\mathcal{N}$ be the smallest σ-algebra of subsets of N such that for each $B \in \mathcal{B}_0(\mathbb{R}^d)$ the mapping that assigns the value $\varphi(B)$ to each measure $\varphi \in \mathsf{N}$ is $(\mathcal{N}, \mathcal{B}(\mathbb{R}))$-measurable.

Definition 3.3.5 (Random counting measure)**.** A *random counting measure* $N : \Omega \to \mathsf{N}$ is a random element defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in the measurable space $(\mathsf{N}, \mathcal{N})$.

Remark 3.3.4. Analogous to random measures, see Section 3.3.1, random counting measures can be interpreted as random functions with index set $\mathcal{T} = \mathcal{B}(\mathbb{R}^d)$. Thus, we often use the notation $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ for a random counting measure N, where $N(B)$ denotes the random value of N for each Borel set $B \in \mathcal{B}(\mathbb{R}^d)$. Accordingly, the distribution of N is uniquely determined by the family of finite dimensional distributions of N, compare to Theorem 3.2.2.

Lemma 3.3.1. *Let $X = \{X_i, i = 1, 2, \dots\}$ be a random point process and define*

$$N(B) = \#\{i : X_i \in B\}, \quad B \in \mathcal{B}(\mathbb{R}^d). \tag{3.19}$$

Then, $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ is a random counting measure. Vice versa, for each random counting measure $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ there is a uniquely determined random point process $X = \{X_i, i = 1, 2, \dots\}$ such that (3.19) *holds.*

Proof. If $X = \{X_i, i = 1, 2, \dots\}$ is a random point process, it can easily be checked that for $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ as defined in (3.19) we have that $N(B) \in \mathbb{N}_0 \cup \{\infty\}$ for all $B \in \mathcal{B}(\mathbb{R}^d)$, $N(\emptyset) = 0$, and that N is σ-additive almost surely. For a proof of the reverse statement see Daley and Vere-Jones (2008), Lemma 9.1.XIII. □

Remark 3.3.5.

1. According to Lemma 3.3.1, each random point process can also be interpreted as a random counting measure and vice versa. Thus, we will use these two notations interchangeably in the following.
2. Since each random counting measure is a special case of a random measure, the definition of the intensity measure of a random counting measure $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ is included in Definition 3.3.2. The intensity measure of a point process $X = \{X_i, i = 1, 2, \dots\}$ is defined as the intensity measure of the corresponding random counting measure N.
3. In the following, we only consider random point processes with locally finite and diffuse intensity measures. A measure $\mu \in \mathsf{M}$ is called diffuse if $\mu(\{x\}) = 0$ for all $x \in \mathbb{R}^d$.
4. We call a random point process $X = \{X_i, i = 1, 2, \dots\}$ stationary if its corresponding random counting measure $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ is stationary in the sense of Definition 3.3.3. The same applies for the notation of isotropy.

In the case of stationarity, the intensity measure of a point process simplifies as follows.

Theorem 3.3.1. *Let $X = \{X_i, i = 1, 2, \dots\}$ be a stationary random point process with intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$. Then, there is a constant $\lambda \in [0, \infty)$ such that*

$$\mu(B) = \lambda\, \nu_d(B), \quad B \in \mathcal{B}(\mathbb{R}^d).$$

The constant λ is called intensity *of X.*

Proof. Let $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ be the corresponding random counting measure of X. From the stationarity of N it follows that

$$\mu(B) = \mathbb{E}\, N(B) = \mathbb{E}\, N(B + x) = \mu(B + x), \quad x \in \mathbb{R}^d, B \in \mathcal{B}(\mathbb{R}^d).$$

Thus, μ is a translation invariant measure, which implies that μ is a multiple of the Lebesgue measure, see Schilling (2005), Theorem 5.8. The intensity λ is finite since the intensity measure μ is locally finite. □

Remark 3.3.6. In applications, when a point process $X = \{X_i, i = 1, 2, \ldots\}$ with intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ cannot be considered to be stationary, it is often assumed that μ is absolutely continuous with respect to the d-dimensional Lebesgue measure. In that case, there is a Borel-measurable, locally integrable function $\lambda : \mathbb{R}^d \to [0, \infty)$, called *intensity function* of X, such that

$$\mu(B) = \int_B \lambda(t)\, \mathrm{d}t, \quad B \in \mathcal{B}(\mathbb{R}^d).$$

The following result is essentially an application of Fubini's theorem, which goes back to the early work of Campbell (Campbell, 1909).

Theorem 3.3.2 (Campbell)**.** *Let $X = \{X_i, i = 1, 2, \ldots\}$ be a random point process with intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$. For each Borel-measurable function $f : \mathbb{R}^d \cup \{\infty\} \to [0, \infty)$ with $f(\infty) = 0$ it holds that*

$$\mathbb{E}\left(\sum_{i=1}^{\infty} f(X_i)\right) = \int_{\mathbb{R}^d} f(t)\, \mu(\mathrm{d}t).$$

Proof. See Schneider and Weil (2008), Theorem 3.1.2.

An important characteristic of random point processes is the probability generating functional, which is a generalization of the probability generating function of nonnegative discrete random variables. Like its counterpart for random variables, the probability generating functional uniquely determines the distribution of a random point process. Let F be the family of all Borel-measurable functions $f : \mathbb{R}^d \cup \{\infty\} \to [0, 1]$ such that $f(x) = 1$ for $x \notin B_f$ with some bounded Borel set $B_f \in \mathcal{B}_0(\mathbb{R}^d)$ (which depends on f).

Definition 3.3.6 (Probability generating functional)**.** Let $X = \{X_i, i = 1, 2, \ldots\}$ be a random point process. The mapping $G : \mathsf{F} \to [0, 1]$ defined by

$$G(f) = \mathbb{E}\left(\prod_{i=1}^{\infty} f(X_i)\right), \quad f \in \mathsf{F},$$

is called *probability generating functional* of X.

Theorem 3.3.3. *Let* $X^{(1)} = \{X_i^{(1)}, i = 1, 2, \ldots\}$ *and* $X^{(2)} = \{X_i^{(2)}, i = 1, 2, \ldots\}$ *be random point processes with probability generating functionals* $G^{(1)} : \mathsf{F} \to [0,1]$ *and* $G^{(2)} : \mathsf{F} \to [0,1]$. *The point processes* $X^{(1)}$ *and* $X^{(2)}$ *have the same distribution if and only if*

$$G^{(1)}(f) = G^{(2)}(f), \quad f \in \mathsf{F}.$$

Proof. See Cressie (1993), Section 8.3.2.

The distributions of random point processes can be further characterized by means of (factorial) moment measures, which represent point process analogues to the (factorial) moments of random variables and random vectors.

Definition 3.3.7 (mth moment measure)**.** Let $X = \{X_i, i = 1, 2, \ldots\}$ be a random point process and $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ the corresponding random counting measure. For each $m \in \mathbb{N}$ with $\mathbb{E}\, N^m(B) < \infty$ for all $B \in \mathcal{B}_0(\mathbb{R}^d)$, the *$m$th moment measure* $\mu_m : \mathcal{B}(\mathbb{R}^{md}) \to [0, \infty]$ of X is defined by

$$\mu_m(B_1 \times \ldots \times B_m) = \mathbb{E}\left(\sum_{(i_1,\ldots,i_m)} \mathbb{1}_{B_1 \times \ldots \times B_m}(X_{i_1}, \ldots, X_{i_m}) \right), \quad B_1, \ldots, B_m \in \mathcal{B}(\mathbb{R}^d), \tag{3.20}$$

where the summation is over all m-tuples $(i_1, \ldots, i_m)$ with $i_1, \ldots, i_m \in \mathbb{N}$.

Definition 3.3.8 (mth factorial moment measure)**.** Let $X = \{X_i, i = 1, 2, \ldots\}$ be a random point process and $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ the corresponding random counting measure. For each $m \in \mathbb{N}$ with $\mathbb{E}\, N^m(B) < \infty$ for all $B \in \mathcal{B}_0(\mathbb{R}^d)$, the *$m$th factorial moment measure* $\alpha_m : \mathcal{B}(\mathbb{R}^{md}) \to [0, \infty]$ of X is defined by

$$\alpha_m(B_1 \times \ldots \times B_m) = \mathbb{E}\left(\sum_{(i_1,\ldots,i_m)}^{\neq} \mathbb{1}_{B_1 \times \ldots \times B_m}(X_{i_1}, \ldots, X_{i_m}) \right), \quad B_1, \ldots, B_m \in \mathcal{B}(\mathbb{R}^d),$$

where the summation is over all m-tuples $(i_1, \ldots, i_m)$ with $i_1, \ldots, i_m \in \mathbb{N}$ and $i_j \neq i_k$ for $j \neq k$.

Remark 3.3.7. Let $X = \{X_i, i = 1, 2, \ldots\}$ be a random point process with mth moment measure $\mu_m : \mathcal{B}(\mathbb{R}^{md}) \to [0, \infty]$ and mth factorial moment measure $\alpha_m : \mathcal{B}(\mathbb{R}^{md}) \to [0, \infty]$ for $m \in \mathbb{N}$.

1. The first moment measure μ_1 and the first factorial moment measure α_1 are both given by the intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ of X. Furthermore, in the case $m = 2$ we get that for arbitrary $B_1, B_2 \in \mathcal{B}(\mathbb{R}^d)$

$$\mu_2(B_1 \times B_2) = \alpha_2(B_1 \times B_2) + \mu(B_1 \cap B_2), \tag{3.21}$$

which can easily be seen if the sum on the right-hand side of (3.20) is split up as follows:

$$\sum_{(i_1,i_2)} \mathbb{1}_{B_1\times B_2}(X_{i_1},X_{i_2}) = \sum_{(i_1,i_2)}^{\neq} \mathbb{1}_{B_1\times B_2}(X_{i_1},X_{i_2}) + \sum_{i=1}^{\infty} \mathbb{1}_{B_1\cap B_2}(X_i).$$

2. The definition of μ_m implies the following generalization of the Campbell formula stated in Theorem 3.3.2. For any nonnegative Borel-measurable function $f : \mathbb{R}^{md} \to [0,\infty)$ it holds that

$$\mathbb{E}\left(\sum_{(i_1,\ldots,i_m)} f(X_{i_1},\ldots,X_{i_m})\right) = \int_{\mathbb{R}^{md}} f(t_1,\ldots,t_m)\,\mu_m(\mathrm{d}(t_1,\ldots,t_m)), \tag{3.22}$$

see, e.g., Schneider and Weil (2008), Theorem 3.1.3.

3.3.3. Poisson point processes

We now introduce some classes of point process that are considered later on in this thesis for modeling purposes and describe their most important properties and characteristics as well as some frequently used simulation algorithms. We start with the Poisson point process, which is not only the most popular model in applications but also provides the basis for the construction of more sophisticated models such as Cox or (Poisson) cluster processes. A more detailed consideration of Poisson processes can be found, e.g., in Kingman (1993) and Streit (2010).

Definition 3.3.9 (Poisson point process)**.** Let $X = \{X_i, i = 1,2,\ldots\}$ be a random point process and $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ the corresponding random counting measure. We call X a *Poisson point process* (or Poisson process for short) with (locally finite and diffuse) intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0,\infty]$ if

1. the random variables $N(B_1), N(B_2), \ldots$ are independent of each other for any pairwise disjoint $B_1, B_2, \ldots \in \mathcal{B}_0(\mathbb{R}^d)$,
2. the random number of points $N(B)$ in any $B \in \mathcal{B}_0(\mathbb{R}^d)$ has a Poisson distribution with parameter $\mu(B)$, i.e.,

$$\mathbb{P}(N(B) = n) = \frac{\mu(B)^n}{n!}\exp\{-\mu(B)\}, \quad n \in \mathbb{N}_0. \tag{3.23}$$

Remark 3.3.8. 1. If $X = \{X_i, i = 1,2,\ldots\}$ is a Poisson process, then the corresponding random counting measure $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ is sometimes denoted as *Poisson counting measure.* However, most researchers use the term Poisson process for both X and N, which is also done in the present thesis.

2. The intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ of a Poisson process $X = \{X_i, i = 1, 2, \ldots\}$ completely determines the distribution of the process. In general, this is not the case for random point processes.

Theorem 3.3.4. *Let $X = \{X_i, i = 1, 2, \ldots\}$ be a Poisson process with intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$.*

(i) The probability generating functional $G : \mathsf{F} \to [0, 1]$ of X is given by

$$G(f) = \exp\left\{ \int_{\mathbb{R}^d} (f(t) - 1)\, \mu(\mathrm{d}t) \right\}, \quad f \in \mathsf{F}. \tag{3.24}$$

(ii) The mth factorial moment measure $\alpha_m : \mathcal{B}(\mathbb{R}^{md}) \to [0, \infty]$ of X is given by

$$\alpha_m(B_1 \times \ldots \times B_m) = \mu(B_1) \cdot \ldots \cdot \mu(B_m), \quad B_1, \ldots, B_m \in \mathcal{B}(\mathbb{R}^d). \tag{3.25}$$

Proof. For *(i)* see Chiu et al. (2013), Example 4.2. and for *(ii)* see Schneider and Weil (2008), Corollary 3.2.4.

Theorem 3.3.5. *Let $X = \{X_i, i = 1, 2, \ldots\}$ be a Poisson process with intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$. For any nonnegative Borel-measurable functions $f, g : \mathbb{R}^d \to [0, \infty)$ it holds that*

$$\mathbb{E}\left(\sum_{i=1}^{\infty} f(X_i) \sum_{j=1}^{\infty} g(X_j) \right) = \int_{\mathbb{R}^d} f(t_1)\, \mu(\mathrm{d}t_1) \int_{\mathbb{R}^d} g(t_2)\, \mu(\mathrm{d}t_2) + \int_{\mathbb{R}^d} f(t)g(t)\, \mu(\mathrm{d}t). \tag{3.26}$$

Proof. The statement follows immediately by using the more general definition of the mth moment measure in (3.22) with $m = 2$, the relationship between the second moment measure, the second factorial moment measure and the intensity measure in (3.21), and the representation formula of the mth factorial moment measure of a Poisson process in (3.25) with $m = 2$ as follows:

$$\begin{aligned}
\mathbb{E}\left(\sum_{(i,j)} f(X_i)\, g(X_j) \right) &= \int_{\mathbb{R}^{2d}} f(t_1)g(t_2)\, \mu_2(\mathrm{d}(t_1, t_2)) \\
&= \int_{\mathbb{R}^{2d}} f(t_1)g(t_2)\, \alpha_2(\mathrm{d}(t_1, t_2)) + \int_{\mathbb{R}^d} f(t)g(t)\, \mu(\mathrm{d}t) \\
&= \int_{\mathbb{R}^d} f(t_1)\, \mu(\mathrm{d}t_1) \int_{\mathbb{R}^d} g(t_2)\, \mu(\mathrm{d}t_2) + \int_{\mathbb{R}^d} f(t)g(t)\, \mu(\mathrm{d}t). \quad \square
\end{aligned}$$

To conclude the discussion of the first point process model, we suggest several simulation algorithms that allow to generate realizations of (stationary and non-stationary) Poisson processes in general compact windows. At first, consider a stationary Poisson process $X = \{X_i, i = 1, 2, \ldots\}$ with intensity $\lambda > 0$ and the corresponding random counting measure $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$. Furthermore, let $W = [a_1, b_1] \times \ldots \times [a_d, b_d]$ with $a_i < b_i$ for $i = 1, \ldots, d$ be a cuboid in $\mathbb{R}^d$, in which a realization of X is to be simulated. Then, according to Definition 3.3.9 and Theorem 3.3.1, the random number $N(W)$ of points in W is Poisson distributed with parameter $\lambda\nu_d(W)$. Furthermore, conditioned on $N(W) = n$ for some $n \in \mathbb{N}_0$, the restriction of X to W consists of n independent random vectors, which are uniformly distributed in W, see, e.g., Illian et al. (2008), Section 2.3.2 (e). This motivates the following algorithm for the simulation of X in W. For the generation of Poisson and standard uniformly distributed pseudo-random numbers and uniformly in W distributed pseudo-random vectors we refer to Gentle (2003) or Kroese et al. (2011).

Algorithm 3.3.1. 1. Generate a Poisson distributed pseudo-random number n with parameter $\lambda\nu_d(W)$.

2. Generate n uniformly distributed pseudo-random vectors $x_1, \ldots, x_n$ in W.

Then, $\{x_1, \ldots, x_n\}$ can be regarded as a realization of a stationary Poisson process with intensity $\lambda > 0$ restricted to W.

In order to simulate Poisson processes under more general conditions, the following theorem concerning location-dependent thinning turns out to be useful. It motivates a class of algorithms that are denoted as *acceptance-rejection methods* in literature.

Theorem 3.3.6. *Let $X = \{X_i, i = 1, 2, \ldots\}$ be a Poisson process with (Borel-measurable and locally integrable) intensity function $\lambda_1 : \mathbb{R}^d \to [0, \infty)$ and let $\lambda_2 : \mathbb{R}^d \to [0, \infty)$ be another Borel-measurable and locally integrable function such that $\lambda_1(x) \geq \lambda_2(x)$ for all $x \in \mathbb{R}^d$. Furthermore, consider a sequence $U_1, U_2, \ldots$ of independent, standard uniformly distributed random variables that are independent of X. Then, the random counting measure $\tilde{N} = \{\tilde{N}(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ defined by*

$$\tilde{N}(B) = \#\left\{i : X_i \in B, U_i \leq \frac{\lambda_2(X_i)}{\lambda_1(X_i)}\right\}, \quad B \in \mathcal{B}(\mathbb{R}^d),$$

is a Poisson process with intensity function λ_2.

Proof. See Møller and Waagepetersen (2004), Proposition 3.7.

Based on this theorem an algorithm for the simulation of a stationary Poisson process $X = \{X_i, i = 1, 2, \ldots\}$ with intensity $\lambda > 0$ in an arbitrary compact Borel set $B \in \mathcal{B}_0(\mathbb{R}^d)$ can be given. For that purpose, let $W = [a_1, b_1] \times \ldots \times [a_d, b_d]$ with $a_i < b_i$

for $i = 1, \ldots, d$ be the smallest d-dimensional cuboid such that $B \subset W$. Applying Theorem 3.3.6 with $\lambda_1(x) = \lambda \, \mathbb{1}_W(x)$ and $\lambda_2(x) = \lambda \, \mathbb{1}_B(x)$ provides the basis for the following simulation algorithm.

Algorithm 3.3.2. 1. Generate a realization $\{x_1, \ldots, x_n\}$ of a stationary Poisson process with intensity $\lambda > 0$ in W according to Algorithm 3.3.1.

2. Eliminate those points of $\{x_1, \ldots, x_n\}$ that are not located in B and denote the set of remaining points as $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ with $m \leq n$.

Then, $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ can be regarded as a realization of a stationary Poisson process with intensity $\lambda > 0$ restricted to B.

Remark 3.3.9. If B is a d-dimensional ball $b(x, r)$ with some center $x \in \mathbb{R}^d$ and a positive radius $r > 0$, then it is more elegant to simulate a stationary Poisson process $X = \{X_i, i = 1, 2, \ldots\}$ in B according to a radial simulation algorithm proposed in Quine and Watson (1984). Additionally, Theorem 3.3.6 can be used to generate a realization of X in a subset $\widetilde{B} \subset B$ using acceptance-rejection, which may be more efficient than simulating X in a d-dimensional cuboid $W \subset \mathbb{R}^d$ first.

Finally, Theorem 3.3.6 is also applied to derive a simulation algorithm for the non-stationary case. Consider a Poisson process $X = \{X_i, i = 1, 2, \ldots\}$ with intensity function $\lambda : \mathbb{R}^d \to [0, \infty)$ and an arbitrary compact Borel set $B \in \mathcal{B}_0(\mathbb{R}^d)$ such that $\lambda_{max} = \max\{\lambda(x), x \in B\} > 0$. Then, using Theorem 3.3.6 with $\lambda_1(x) = \lambda_{max} \, \mathbb{1}_B(x)$ and $\lambda_2(x) = \lambda(x) \, \mathbb{1}_B(x)$ for all $x \in \mathbb{R}^d$ motivates the following simulation algorithm.

Algorithm 3.3.3. 1. Generate a realization $\{x_1, \ldots, x_n\}$ of a stationary Poisson process with intensity $\lambda_{max} > 0$ in B according to Algorithm 3.3.2.

2. For each $i \in \{1, \ldots, n\}$ generate a standard uniformly distributed pseudo-random number u_i. If $u_i \leq \lambda(x_i)/\lambda_{max}$, then accept x_i, otherwise reject x_i. Denote the set of accepted points as $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ with $m \leq n$.

Then, $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ can be regarded as a realization of a (non-stationary) Poisson process with intensity function $\lambda : \mathbb{R}^d \to [0, \infty)$ restricted to B.

3.3.4. Cox point processes

An intuitive way to construct a more flexible model for random point processes is to randomize the intensity measure of a Poisson process, which leads to the class of Cox point processes (also called doubly stochastic Poisson processes in literature). Cox processes were first studied systematically in Cox (1955).

Definition 3.3.10 (Cox point process)**.** Let $X = \{X_i, i = 1, 2, \ldots\}$ be a random point process, $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ the corresponding random counting measure, and $M = \{M(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ a random measure, which is locally finite and diffuse almost surely. We call X a *Cox point process* (or Cox process for short) with random intensity measure M if

$$\mathbb{P}\left(\bigcap_{i=1}^{n}\{N(B_i) = k_i\}\right) = \mathbb{E}\left(\prod_{i=1}^{n}\frac{M(B_i)^{k_i}}{k_i!}\exp\{-M(B_i)\}\right)$$

for all $n \in \mathbb{N}$, $k_1, \ldots, k_n \in \mathbb{N}_0$ and pairwise disjoint $B_1, \ldots, B_n \in \mathcal{B}_0(\mathbb{R}^d)$.

Remark 3.3.10. Let $X = \{X_i, i = 1, 2, \ldots\}$ be a Cox process with random intensity measure $M = \{M(B), B \in \mathcal{B}(\mathbb{R}^d)\}$.

1. According to Definition 3.3.10, the distribution of X can be interpreted as a mixture of the distributions of different Poisson processes. In particular, conditioned on $\{M = \mu\}$ for any (locally finite and diffuse) measure $\mu \in \mathsf{M}$, the process X is a Poisson process with intensity measure μ.

2. The intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ of X coincides with the intensity measure of M, i.e., $\mu(B) = \mathbb{E}\, M(B)$ for all $B \in \mathcal{B}(\mathbb{R}^d)$, see, e.g., Schneider and Weil (2008), Section 3.2.

3. A special case, which is particularly important in applications, is given, when the random intensity measure M is absolutely continuous with respect to the d-dimensional Lebesgue measure almost surely. This implies that there exists a random field $\Lambda = \{\Lambda(t), t \in \mathbb{R}^d\}$ with $\Lambda(t) : \Omega \to [0, \infty)$ for $t \in \mathbb{R}^d$, which has Borel-measurable and locally integrable realizations, such that

$$M(B) = \int_B \Lambda(t)\,\mathrm{d}t, \quad B \in \mathcal{B}(\mathbb{R}^d),$$

almost surely. The random field Λ is denoted as *random intensity function* of X.

3.3.5. Cluster processes

As a third class of random point processes that are used for modeling purposes in the present thesis random cluster processes are introduced. We give a general definition of cluster processes and describe commonly used models that are constructed based on independent Poisson processes. In accordance to the applications discussed later on in this thesis, we only consider processes which have a finite number of clusters almost surely.

Definition 3.3.11 (Cluster process)**.** Let $X^{(0)} = \{X_i^{(0)}, i = 1, \ldots, Z\}$ be a random point process with finite and diffuse intensity measure $\mu^{(0)} : \mathcal{B}(\mathbb{R}^d) \to [0, \infty)$, where $Z : \Omega \to \mathbb{N}_0$ is a random variable describing the total number of points of $X^{(0)}$ in $\mathbb{R}^d$. Furthermore, consider a sequence of identically distributed random counting measures $N^{(1)} = \{N^{(1)}(B), B \in \mathcal{B}(\mathbb{R}^d)\}, N^{(2)} = \{N^{(2)}(B), B \in \mathcal{B}(\mathbb{R}^d)\}, \ldots$ with finite and diffuse intensity measure $\mu^{(1)} : \mathcal{B}(\mathbb{R}^d) \to [0, \infty)$. If the random counting measure $N = \{N(B), B \in \mathcal{B}(\mathbb{R}^d)\}$ defined by

$$N(B) = \sum_{i=1}^{Z} N^{(i)}(B - X_i^{(0)}), \quad B \in \mathcal{B}(\mathbb{R}^d),$$

is locally finite and simple almost surely, i.e., if N is a random counting measure in the sense of Definition 3.3.5, then we call N (or the corresponding random point process $X = \{X_i, i = 1, 2, \ldots\}$) a *cluster process*.

Remark 3.3.11. Let $X = \{X_i, i = 1, 2, \ldots\}$ be a cluster process as introduced in Definition 3.3.11.

1. We call the random point process $X^{(0)}$ the *primary process* and the random counting measures $N^{(1)}, N^{(2)}, \ldots$ the *secondary processes* of X. If the primary process is a Poisson process, then X is called a *Poisson cluster process*.

2. In most applications it is assumed that the secondary processes $N^{(1)}, N^{(2)}, \ldots$ are independent of each other and of the primary process $X^{(0)}$. In this case, the (finite and diffuse) intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty)$ of X is given by

$$\mu(B) = \int_{\mathbb{R}^d} \mu^{(1)}(B - t)\, \mu^{(0)}(\mathrm{d}t), \quad B \in \mathcal{B}(\mathbb{R}^d),$$

   see, e.g., Daley and Vere-Jones (2003), (6.3.3).

We introduce a subclass of Poisson cluster processes, called Neyman-Scott processes, which can also be interpreted as special Cox processes. The name of the Neyman-Scott process refers to its occurrence in Neyman and Scott (1958). In the following, we suppose that the primary process and the secondary processes have intensity functions (i.e., their intensity measures are absolutely continuous with respect to the d-dimensional Lebesgue measure), which is a common assumption in applications.

Example 3.3.1 (Neyman-Scott process)**.** Consider a Poisson cluster process $X = \{X_i, i = 1, 2, \ldots\}$, where the primary process $X^{(0)} = \{X_i^{(0)}, i = 1, \ldots, Z\}$ is a Poisson process with integrable intensity function $\lambda^{(0)} : \mathbb{R}^d \to [0, \infty)$ and the secondary processes $N^{(1)} = \{N^{(1)}(B), B \in \mathcal{B}(\mathbb{R}^d)\}, N^{(2)} = \{N^{(2)}(B), B \in \mathcal{B}(\mathbb{R}^d)\}, \ldots$ are independent and identically distributed Poisson processes with integrable intensity function $\lambda^{(1)} : \mathbb{R}^d \to$

$[0,\infty)$, which are independent of $X^{(0)}$. Then, X is called a *Neyman-Scott process* and the (finite) intensity measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0,\infty)$ of X is given by

$$\mu(B) = \int_{\mathbb{R}^d} \int_B \lambda^{(1)}(s-t)\, \mathrm{d}s\, \lambda^{(0)}(t)\, \mathrm{d}t, \quad B \in \mathcal{B}(\mathbb{R}^d).$$

One can show that X is also a Cox process with random intensity function $\Lambda = \{\Lambda(t), t \in \mathbb{R}^d\}$ given by

$$\Lambda(t) = \sum_{i=1}^{Z} \lambda^{(1)}(t - X_i^{(0)}), \quad t \in \mathbb{R}^d,$$

see, e.g., Cressie (1993), Section 8.5.3.

When the intensity function of the secondary processes simplifies to being constant in a ball around the origin, i.e., when $\lambda^{(1)}(t) = \lambda^{(1)} \mathbb{1}_{b(o,r)}(t)$ for $t \in \mathbb{R}^d$ with $\lambda^{(1)}, r > 0$, then we call X a *Matérn cluster process* with intensity function $\lambda^{(0)} : \mathbb{R}^d \to [0,\infty)$ of the primary process, cluster intensity $\lambda^{(1)} > 0$, and cluster radius $r > 0$.

To conclude the discussion of models for random point processes, we propose an algorithm for the simulation of a Matérn cluster process in a compact Borel set $B \in \mathcal{B}_0(\mathbb{R}^d)$. Note that this algorithm does not account for boundary effects and is thus only suitable if the cluster radius r is small compared to the size of B. In order to perform a boundary correction, the values of the intensity function of the primary process need to be known for the extended set $B \oplus b(o,r)$, which is not the case in the applications considered in this thesis. Thus, point patterns generated by Algorithm 3.3.4 are only denoted as approximate realizations in the following.

Algorithm 3.3.4.

1. Generate a realization $\{x_1^{(0)}, \ldots, x_n^{(0)}\}$ of a Poisson process with intensity function $\lambda^{(0)} : \mathbb{R}^d \to [0,\infty)$ in B according to Algorithm 3.3.3.
2. For each $i \in \{1,\ldots,n\}$, generate a realization $\{x_1^{(i)}, \ldots, x_{n_i}^{(i)}\}$ of a stationary Poisson process with intensity $\lambda^{(1)} > 0$ in $b(x_i^{(0)}, r) \cap B$ according to Algorithm 3.3.2.

Then, the union $\bigcup_{i=1}^n \{x_1^{(i)}, \ldots, x_{n_i}^{(i)}\}$ can be regarded as an approximate realization of a Matérn cluster process with intensity function $\lambda^{(0)} : \mathbb{R}^d \to [0,\infty)$ of the primary process, cluster intensity $\lambda^{(1)} > 0$, and cluster radius $r > 0$ restricted to B.

3.3.6. Random closed sets and germ-grain models

In many applications objects need to be modeled that have a different (and more complicated) geometric structure than a point pattern. For that purpose, we discuss a class of random elements, called random closed sets, that serve as a general tool for the modeling of irregular geometric objects. We introduce a special subclass of random closed sets that are defined based on random point processes and describe suitable simulation algorithms. Besides the references given at the beginning of Section 3.3, we recommend Molchanov (2005) and Nguyen (2006) for a more detailed discussion of this topic. Let C denote the family of all closed sets in $\mathbb{R}^d$ and K be the family of all compact sets in $\mathbb{R}^d$. Furthermore, by $\mathcal{C}$ we denote the smallest σ-algebra of subsets of C that contains the sets $\{C \in \mathsf{C} : C \cap K \neq \emptyset\}$ for all $K \in \mathsf{K}$.

Definition 3.3.12 (Random closed set)**.** A *random closed set* $\Xi : \Omega \to \mathsf{C}$ is a random element defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in the measurable space $(\mathsf{C}, \mathcal{C})$.

Definition 3.3.13 (Stationarity and isotropy)**.** A random closed set Ξ is said to be

1. *stationary* if $\Xi \stackrel{\mathrm{d}}{=} \Xi + x$ for any translation vector $x \in \mathbb{R}^d$,
2. *isotropic* if $\Xi \stackrel{\mathrm{d}}{=} \delta\Xi$ for any rotation $\delta : \mathbb{R}^d \to \mathbb{R}^d$ around the origin.

In general, it is difficult, if not impossible, to give a closed representation formula for the distribution P_Ξ of a random closed set Ξ. A much more elegant way to characterize the distribution of Ξ is using the so-called capacity functional.

Definition 3.3.14 (Capacity functional)**.** Let Ξ be a random closed set in $\mathbb{R}^d$. We call the function $T_\Xi : \mathsf{K} \to [0, 1]$ with

$$T_\Xi(K) = P(\Xi \cap K \neq \emptyset), \quad K \in \mathsf{K},$$

the *capacity functional* of Ξ.

Theorem 3.3.7. *Let Ξ_1 and Ξ_2 be random closed sets in $\mathbb{R}^d$ with capacity functionals $T_{\Xi_1} : \mathsf{K} \to [0, 1]$ and $T_{\Xi_2} : \mathsf{K} \to [0, 1]$. Then, Ξ_1 and Ξ_2 have the same distribution if and only if $T_{\Xi_1}(K) = T_{\Xi_2}(K)$ for all $K \in \mathsf{K}$.*

Proof. It follows directly from the definition of the capacity functional that T_{Ξ_1} and T_{Ξ_2} are identical if $\Xi_1 \stackrel{\mathrm{d}}{=} \Xi_2$. For the converse statement see Schneider and Weil (2008), Theorem 2.1.3. □

As a special case of random closed sets we consider germ-grain models, which are closely related to random point processes. To be more precise, germ-grain models are constructed by assigning a bounded random closed set to each point of a point process and by taking the union of the obtained sets. In accordance to the applications discussed later on in this thesis, we only consider germ-grain models that are compact almost surely in the following.

Definition 3.3.15 (Germ-grain model)**.** Let $X = \{X_i, i = 1, \dots, Z\}$ be a random point process in $\mathbb{R}^d$, where $Z : \Omega \to \mathbb{N}_0$ is a random variable describing the total number of points of X in $\mathbb{R}^d$ with $\mathbb{P}(Z < \infty) = 1$. Furthermore, let $\Xi_1, \Xi_2, \dots$ be a sequence of identically distributed random closed sets in $\mathbb{R}^d$, which are bounded almost surely. Then, the random union set

$$\Xi = \bigcup_{i=1}^{Z} (X_i + \Xi_i)$$

is a bounded random closed set almost surely, which is called a *germ-grain model*. The random vectors $X_1, \dots, X_Z$ are denoted as *germs* and the random closed sets $\Xi_1, \dots, \Xi_Z$ as *grains*.

We present an example of a germ-grain model, which is one of the most commonly used models in stochastic geometry. In the following, we assume that the random point process describing the germs has an intensity function, i.e., its intensity measure is absolutely continuous with respect to the d-dimensional Lebesgue measure.

Example 3.3.2 (Boolean model)**.** Consider a germ-grain model Ξ, where the random point process $X = \{X_i, i = 1, \dots, Z\}$ of germs is a Poisson process with integrable intensity function $\lambda^{(0)} : \mathbb{R}^d \to [0, \infty)$ and the grains $\Xi_1, \Xi_2, \dots$ are independent of each other and of X. Then, Ξ is called a *Boolean model* with intensity function $\lambda^{(0)}$ and grain distribution P_{Ξ_1}.

In many applications it is assumed that $\Xi_i = b(o, R_i)$ for all $i \in \mathbb{N}$, where $R_1, R_2, \dots : \Omega \to [0, \infty)$ is a sequence of independent and identically distributed random variables with distribution function $F : \mathbb{R} \to [0, 1]$. In this case, we call Ξ a Boolean model with spherical grains. Further simplifications that are often considered in applications include that $R_i = R$ for some random variable $R : \Omega \to [0, \infty)$ or that $R_i = r$ for some deterministic radius $r > 0$.

To conclude this section, we provide algorithms for the simulation of certain germ-grain models in a compact Borel set $B \in \mathcal{B}_0(\mathbb{R}^d)$. Similar to Algorithm 3.3.4, we do not take boundary effects into account. For a boundary correction the intensity function of the random point process describing the germs needs to be known outside of B, which is not the case in the applications considered in this thesis. Thus, sets generated by Algorithms 3.3.5 and 3.3.6 are only denoted as approximate realizations in the following. At first, consider a Boolean model Ξ with spherical grains, which is characterized by the intensity function $\lambda^{(0)} : \mathbb{R}^d \to [0, \infty)$ and the distribution function F of the grain radii. We suppose that an algorithm for the generation of pseudo-random numbers with distribution function F is available. For the most commonly used distributions simulation algorithms can be found in Kroese et al. (2011).

Algorithm 3.3.5. 1. Generate a realization $\{x_1, \ldots, x_n\}$ of a Poisson process with intensity function $\lambda^{(0)} : \mathbb{R}^d \to [0, \infty)$ in B according to Algorithm 3.3.3.

2. Generate n pseudo-random numbers $r_1, \ldots, r_n$ with distribution function F.

Then, the union set $\bigcup_{i=1}^{n}(b(x_i, r_i) \cap B)$ can be regarded as an approximate realization of the Boolean model Ξ as described above restricted to B.

Finally, we provide a simulation algorithm for germ-grain models, whose germs are given by a more general random point process. Consider a germ-grain model Ξ with spherical grains $\Xi_1, \Xi_2, \ldots$, i.e., $\Xi_i = b(o, R_i)$ for all $i \in \mathbb{N}$, where $R_1, R_2, \ldots : \Omega \to [0, \infty)$ is a sequence of independent and identically distributed random variables with distribution function $F : \mathbb{R} \to [0, 1]$. Furthermore, the germs are given by a Matérn cluster process $X = \{X_i, i = 1, \ldots, Z\}$ with intensity function $\lambda^{(0)} : \mathbb{R}^d \to [0, \infty)$ of the primary process, cluster intensity $\lambda^{(1)} > 0$, and cluster radius $r > 0$, where the germs are considered to be independent of the grains.

Algorithm 3.3.6. 1. Generate an approximate realization $\{x_1, \ldots, x_n\}$ of a Matérn cluster process with intensity function $\lambda^{(0)} : \mathbb{R}^d \to [0, \infty)$ of the primary process, cluster intensity $\lambda^{(1)} > 0$, and cluster radius $r > 0$ in B according to Algorithm 3.3.4.

2. Generate n pseudo-random numbers $r_1, \ldots, r_n$ with distribution function F.

Then, the union set $\bigcup_{i=1}^{n}(b(x_i, r_i) \cap B)$ can be regarded as an approximate realization of the germ-grain model Ξ as described above restricted to B.

3.4. Multivariate kernel smoothing

Smoothing denotes a broad class of nonparametric statistical methods for the estimation of high-dimensional unknown functions such as, e.g., probability density functions (PDFs), regression functions or conditional quantiles based on observed data. According to Härdle et al. (2004), the role of smoothing is to "extract structural elements of variable complexity from patterns of random variation" and it is "designed to simultaneously estimate and model the underlying structure". In the present thesis, we focus on kernel smoothing methods, which provide particularly intuitive and simple ways of finding structure in datasets without imposing parametric (or semiparametric) models. Other smoothing methods such as spline smoothing, median smoothing, wavelets or (generalized) additive models are discussed, e.g., in Schimek (2000) or Härdle et al. (2004). While parametric statistical models and methods are widely recognized to be more powerful but require some prior knowledge of the considered structure, their nonparametric counterparts are much more flexible and broadly applicable but may suffer a loss of efficiency (Scott, 1992). Especially in a high-dimensional context often only few parametric models exist and prior information on the underlying structure is more difficult to derive, which can result in incorrectly specified parametric models. In such a case, Scott (1992) recommends to accept the lower efficiency of nonparametric models rather than taking the risk of introducing large biases due to incorrect model specification that cannot be eliminated by increased sample sizes alone. In this section, we address two of the most popular smoothing methods: kernel density estimation and kernel regression. For a more detailed discussion of this topic we recommend Nadaraya (1989), Härdle (1991), Scott (1992), Wand and Jones (1995), Simonoff (1996), and Härdle et al. (2004).

3.4.1. Kernel functions

As the name suggests, methods used in kernel smoothing typically involve the application of so-called kernel functions.

Definition 3.4.1 (Kernel function)**.** A nonnegative, bounded and Borel-measurable function $\kappa : \mathbb{R}^d \to [0, \infty)$ is called a *kernel function* (or *kernel* for short) if

$$\int_{\mathbb{R}^d} \kappa(x) \, \mathrm{d}x = 1, \tag{3.27}$$

$$\int_{\mathbb{R}^d} x\kappa(x) \, \mathrm{d}x = o, \tag{3.28}$$

and

$$\int_{\mathbb{R}^d} xx^\top \kappa(x) \, \mathrm{d}x = \mu_2(\kappa) \, I, \tag{3.29}$$

where $\mu_2(\kappa) \in (0, \infty)$ with

$$\mu_2(\kappa) = \int_{\mathbb{R}^d} x_i^2 \kappa(x_1, \ldots, x_d) \, \mathrm{d}(x_1, \ldots, x_d)$$

is independent of $i \in \{1, \ldots, d\}$. Note that (3.28) and (3.29) are matrix equations.

Remark 3.4.1. In most of the scientific literature a bounded, Borel-measurable function $\kappa : \mathbb{R}^d \to [0, \infty)$ is only required to satisfy (3.27) in order to be called a kernel function. However, on the one hand, (3.28) and (3.29) need to be fulfilled to allow for a derivation of (approximate) statistical properties of the estimators introduced in Sections 3.4.2 and 3.4.3. On the other hand, there are hardly any kernel functions used in applications that do not satisfy (3.28) or (3.29). Therefore, it seems reasonable to only consider kernel functions with the properties specified in Definition 3.4.1.

Example 3.4.1 (One-dimensional kernel functions)**.** There is a wide variety of one-dimensional kernel functions suggested in literature. In particular, any bounded, symmetric PDF of a random variable with expectation zero and finite second moment is a valid kernel function. For example, the function $\kappa : \mathbb{R} \to [0, \infty)$ is called a one-dimensional

1. *uniform kernel* if
$$\kappa(x) = \frac{1}{2}\mathbb{1}_{[-1,1]}(x), \quad x \in \mathbb{R},$$
2. *triangle kernel* if
$$\kappa(x) = (1 - |x|)\mathbb{1}_{[-1,1]}(x), \quad x \in \mathbb{R},$$
3. *Epanechnikov kernel* if
$$\kappa(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{[-1,1]}(x), \quad x \in \mathbb{R}, \tag{3.30}$$
4. *biweight kernel* or *quartic kernel* if
$$\kappa(x) = \frac{15}{16}(1 - x^2)^2\mathbb{1}_{[-1,1]}(x), \quad x \in \mathbb{R},$$
5. *triweight kernel* if
$$\kappa(x) = \frac{35}{32}(1 - x^2)^3\mathbb{1}_{[-1,1]}(x), \quad x \in \mathbb{R},$$
6. *Gaussian kernel* if
$$\kappa(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}, \quad x \in \mathbb{R}. \tag{3.31}$$

Example 3.4.2 (d-dimensional kernel functions)**.** When considering applications in arbitrary dimension $d \geq 2$, it is common to use kernel functions that are constructed based on one-dimensional kernels. Two approaches are suggested for that purpose, see, e.g., Wand and Jones (1995), Section 4.2. Let $\kappa : \mathbb{R} \to [0, \infty)$ be a symmetric one-dimensional kernel function.

1. The function $\kappa^P : \mathbb{R}^d \to [0, \infty)$ defined by
$$\kappa^P(x) = \prod_{i=1}^{d} \kappa(x_i), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d,$$
is called a d-dimensional *product kernel*. One can easily see that κ^P satisfies (3.27)-(3.29) due to κ being a valid kernel function.

2. The function $\kappa^S : \mathbb{R}^d \to [0, \infty)$ defined by
$$\kappa^S(x) = \frac{1}{c_d}\kappa\left(\sqrt{x^\top x}\right), \quad x \in \mathbb{R}^d,$$
with
$$c_d = \int_{\mathbb{R}^d} \kappa\left(\sqrt{x^\top x}\right) \,\mathrm{d}x$$
is called a d-dimensional *radially symmetric kernel*. According to Wand (1992), Section 2.1, κ^S indeed is a kernel function in the sense of Definition 3.4.1.

In general, the product kernel κ^P and the radially symmetric kernel κ^S are different. An exception is given by the d-dimensional Gaussian kernel. If κ is the PDF of the standard normal distribution, then $\kappa^P(x) = \kappa^S(x)$ for all $x \in \mathbb{R}^d$.

3.4.2. Multivariate kernel density estimation

An important characteristic describing the distribution of an absolutely continuous random vector is its (multivariate) PDF. Accordingly, density estimation is one of the most fundamental statistical problems. Given data $x_1, \ldots, x_n \in \mathbb{R}^d$ for $n \in \mathbb{N}$ that can be assumed to be realizations of some independent and identically distributed absolutely continuous random vectors $X_1, \ldots, X_n$ with (unknown) PDF $f : \mathbb{R}^d \to [0, \infty)$, it often is of particular interest in applications to determine values $f(x)$ for certain locations $x \in \mathbb{R}^d$ based on $x_1, \ldots, x_n$. If $d = 1$, the most popular approach is to assume that f belongs to a parametric family $\{f_\theta, \theta \in \Theta\}$ of PDFs with some parameter space $\Theta \subset \mathbb{R}^m$ and $m \in \mathbb{N}$ and to provide an estimator $\hat{\theta}$ of the parameter vector θ based on $X_1, \ldots, X_n$. However, for $d \geq 2$ only few parametric families of PDFs exist, such as the multivariate normal or the multivariate Student distribution, and it is significantly more complicated to find one that fits the underlying data sufficiently well. Also the application of copulas is not always suitable since the marginal distributions of $X_1, \ldots, X_n$ and the type of the copula need to be determined first, which can be a difficult task. In this context, the technique of kernel density estimation serves as an invaluable alternative. Besides the literature given at the beginning of Section 3.4, we recommend Silverman (1986) for more details. In the following, let H be the family of all symmetric and positive definite matrices in $\mathbb{R}^{d\times d}$.

Definition 3.4.2 (Kernel density estimator)**.** Let $X_1, \ldots, X_n$ with $n \in \mathbb{N}$ be a sequence of independent and identically distributed absolutely continuous random vectors in $\mathbb{R}^d$ with PDF $f : \mathbb{R}^d \to [0, \infty)$. Furthermore, let $H \in \mathsf{H}$ be a symmetric and positive definite matrix (called the *bandwidth matrix*) and $\kappa : \mathbb{R}^d \to [0, \infty)$ a d-dimensional kernel function. The estimator $\hat{f}_K$ defined by
$$\hat{f}_K(x) = \frac{1}{n\sqrt{\det(H)}} \sum_{i=1}^{n} \kappa\left(H^{-\frac{1}{2}}(x - X_i)\right), \quad x \in \mathbb{R}^d, \tag{3.32}$$

is called *kernel density estimator* (KDE) of f.

Remark 3.4.2 (Effect of smoothing parametrization)**.** Let $X_1, \ldots, X_n$ and $f : \mathbb{R}^d \to [0, \infty)$ be given as in Definition 3.4.2 and let $\hat{f}_K$ be the KDE introduced in (3.32). The bandwidth matrix H has $\frac{1}{2}d(d+1)$ independent entries, which have to be chosen either manually or automatically. Especially in high-dimensional applications this is a complicated task, which is why H is often assumed to have a simplified structure. For example, sometimes the bandwidth matrix is considered to be a diagonal matrix $H = \mathrm{diag}(h_1^2, \ldots, h_d^2)$ with bandwidths $h_1, \ldots, h_d > 0$. In this case, the KDE $\hat{f}_K$ can be rewritten as

$$\hat{f}_K(x) = \frac{1}{n\, h_1 \ldots h_d} \sum_{i=1}^{n} \kappa \left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_d - X_{id}}{h_d} \right), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d,$$

where $X_i = (X_{i1}, \ldots, X_{id})$ for $i = 1, \ldots, n$. An even more special case is given if $h_i = h$ for $i = 1, \ldots, d$ and a bandwidth $h > 0$, which implies that

$$\hat{f}_K(x) = \frac{1}{n\, h^d} \sum_{i=1}^{n} \kappa \left(\frac{x - X_i}{h} \right), \quad x \in \mathbb{R}^d.$$

Such simplifications of the bandwidth matrix can be interpreted as follows. Assume $\kappa : \mathbb{R}^d \to [0, \infty)$ to be a radially symmetric kernel function with support $b(o, 1)$ (e.g., the d-dimensional radially symmetric Epanechnikov kernel). Then, if $H = h^2 I$ for some $h > 0$, the scaled kernel function $\kappa_H : \mathbb{R}^d \to [0, \infty)$ defined by

$$\kappa_H(x) = \frac{1}{\sqrt{\det(H)}}\, \kappa \left(H^{-\frac{1}{2}} x \right), \quad x \in \mathbb{R}^d, \tag{3.33}$$

is still radially symmetric with support $b(o, h)$. Consequently, the amount of smoothing is the same in every direction and only one bandwidth needs to be determined. Larger values of h expand the support of κ_H and lead to smoother estimates, whereas smaller bandwidths result in estimates showing more local variation. If $H = \mathrm{diag}(h_1^2, \ldots, h_d^2)$ for some $h_1, \ldots, h_d > 0$, then the support of κ_H is a d-dimensional ellipsoid centered at the origin o, whose axes correspond to the coordinate directions. This allows smoothing to be different for all d coordinate directions but at the cost of introducing $d - 1$ additional bandwidths. In order to provide a maximum of flexibility (but also the largest number of smoothing parameters that need to be determined), the bandwidth matrix $H \in \mathsf{H}$ should be chosen to be non-diagonal. In this case, the support of the scaled kernel function κ_H is a d-dimensional ellipsoid centered at o whose axes do not correspond to the coordinate directions, i.e., smoothing is even possible in directions that are different from the coordinate axes. Some simple approaches for the (automatic) selection of the bandwidth matrix are discussed in Section 3.4.4.

Remark 3.4.3. Let $X_1, \ldots, X_n$ and $f : \mathbb{R}^d \to [0, \infty)$ be given as in Definition 3.4.2 and let $\hat{f}_K$ be the KDE introduced in (3.32) with kernel function $\kappa : \mathbb{R}^d \to [0, \infty)$ and bandwidth matrix $H \in \mathsf{H}$.

1. As a direct consequence of (3.27) we get that
$$\mathbb{P}\left(\int_{\mathbb{R}^d} \hat{f}_K(x)\,\mathrm{d}x = 1\right) = 1,$$
i.e., $\hat{f}_K$ is a PDF almost surely.

2. The KDE $\hat{f}_K$ can also be used for a more general purpose than the estimation of a PDF. For example, in stochastic geometry $n\hat{f}_K$ is used as a nonparametric estimator for the intensity function $\lambda : \mathbb{R}^d \to [0, \infty)$ of a random point process X, see, e.g., Illian et al. (2008), Section 3.3 or Diggle (2014), Section 5.3. Scaling $\hat{f}_K$ by the factor n is necessary since for any bounded Borel set $B \in \mathcal{B}_0(\mathbb{R}^d)$ the integral $\int_B \lambda(x)\,\mathrm{d}x$ describes the expected number of points of the process X in B in contrast to the integral $\int_B f(x)\,\mathrm{d}x$, which provides the probability $\mathbb{P}(X_1 \in B)$.

3. Let κ be a radially symmetric kernel defined by $\kappa(x) = c_d^{-1}\kappa_1(\sqrt{x^\top x})$ for $x \in \mathbb{R}^d$ with a normalizing constant $c_d > 0$ and a one-dimensional kernel function $\kappa_1 : \mathbb{R} \to [0, \infty)$ and let H be given as $H = h^2 I$ for some $h > 0$. Then, for each $i \in \{1, \ldots, n\}$ it holds that
$$\kappa\left(H^{-\frac{1}{2}}(x - X_i)\right) = \frac{1}{c_d}\kappa_1\left(\frac{\|x - X_i\|_d}{h}\right), \quad x \in \mathbb{R}^d,$$
i.e., the KDE $\hat{f}_K$ involves the computation of d-dimensional Euclidean distances. However, if $d = 2$ and instead of $\mathbb{R}^2$ a geographical domain $\mathcal{T}$ is considered, which sometimes is the case in geostatistical applications, then it is more suitable to use the great-circle distance $d_{GC}(x, X_i)$ introduced in (3.4) rather than the 2-dimensional Euclidean distance $\|x - X_i\|_2$, see also Remark 3.2.4. Note that this also requires an appropriate modification of the normalizing constant c_2. Utilization of the great-circle distance is rarely needed in density estimation but can be of great importance when estimating intensity functions, compare to the previous point.

We briefly discuss some statistical properties of the KDE such as bias, variance, and mean integrated squared error (MISE). However, it turns out to be very difficult to provide exact representation formulas for these properties under general conditions. Thus, approximate formulas are usually derived by using the multivariate Taylor theorem. In the following, we assume that the bandwidth matrix of the KDE is a multiple of the d-dimensional unit matrix, for the more general case we refer to Wand and Jones (1995), Section 4.3 or Härdle et al. (2004), Section 3.6.

Theorem 3.4.1. *Let $X_1, \ldots, X_n$ with $n \in \mathbb{N}$ be a sequence of independent and identically distributed absolutely continuous random vectors in $\mathbb{R}^d$ with PDF $f : \mathbb{R}^d \to [0, \infty)$, which is assumed to have bounded, continuous, and square integrable second-order partial derivatives. Furthermore, let $\hat{f}_K$ be the KDE of f based on $X_1, \ldots, X_n$ with kernel function $\kappa : \mathbb{R}^d \to [0, \infty)$ and bandwidth matrix $H = h^2 I$ for some $h > 0$. If n is large, h is small and nh^d is large, then*

(i) the expectation of $\hat{f}_K(x)$ can be approximated by

$$\mathbb{E}\, \hat{f}_K(x) \approx f(x) + \frac{1}{2} h^2 \mu_2(\kappa)\, \mathrm{tr}(\mathcal{H}_f(x)), \quad x \in \mathbb{R}^d,$$

(ii) the variance of $\hat{f}_K(x)$ can be approximated by

$$var\, \hat{f}_K(x) \approx \frac{1}{nh^d} f(x) \int_{\mathbb{R}^d} \kappa^2(y)\, \mathrm{d}y, \quad x \in \mathbb{R}^d,$$

(iii) the MISE of $\hat{f}_K$ can be approximated by

$$\mathbb{E}\left(\int_{\mathbb{R}^d} \left(\hat{f}_K(x) - f(x) \right)^2 \mathrm{d}x \right) \approx \frac{1}{nh^d} \int_{\mathbb{R}^d} \kappa^2(y)\, \mathrm{d}y + \frac{1}{4} h^4 \mu_2^2(\kappa) \int_{\mathbb{R}^d} \mathrm{tr}^2(\mathcal{H}_f(y))\, \mathrm{d}y,$$

where $\mathcal{H}_f(x)$ is the Hessian matrix of f at $x \in \mathbb{R}^d$ introduced in Section 3.1.1.

Proof. For a proof in a more general context see Wand (1992), Section 2.1.

Remark 3.4.4. Theorem 3.4.1 reveals a fundamental problem arising when selecting the bandwidth $h > 0$. The (approximate) bias of the estimator $\hat{f}_K(x)$ for any $x \in \mathbb{R}^d$ can be reduced to an arbitrary value by choosing h to be sufficiently small. However, this usually causes the (approximate) variance of $\hat{f}_K(x)$ to be large. On the other hand, one can reduce the variance of $\hat{f}_K(x)$ considerably by arbitrarily increasing h, which comes at the cost of introducing a higher systematic bias into estimation. A good compromise is to determine the smoothing parameter h in such a way that it minimizes the (approximate) MISE, see Section 3.4.4.

Although the KDE is the most popular tool in density estimation, there are applications where it might not be an optimal choice. For example, a particular flaw is that the (constant) bandwidth matrix H is not able to provide different smoothing parameterizations across $\mathbb{R}^d$. While in regions with a large amount of data one is often interested in getting a very detailed picture of the underlying PDF (using rather small bandwidths), one might wish to increase smoothing in regions of sparse data, e.g., in the tails of the underlying PDF (Silverman, 1986, Section 5.1). Such a dynamization of the bandwidth matrix H is denoted as *adaptive smoothing* in literature, see Silverman (1986), Section 5.2-5.3 and Scott (1992), Section 6.6. Two popular yet simple approaches are introduced in the following.

Definition 3.4.3 (Generalized nearest neighbor estimator/variable kernel estimator)**.** Let $X_1, \ldots, X_n$ with $n \in \mathbb{N}$ be a sequence of independent and identically distributed absolutely continuous random vectors in $\mathbb{R}^d$ with PDF $f : \mathbb{R}^d \to [0, \infty)$ and let $\kappa : \mathbb{R}^d \to [0, \infty)$ be a d-dimensional kernel function. Furthermore, $r_k(x)$ denotes the (random) distance of $x \in \mathbb{R}^d$ to the kth nearest of the random vectors $X_1, \ldots, X_n$ for a fixed $k \in \{2, \ldots, n\}$.

1. The estimator $\hat{f}_N$ defined by

$$\hat{f}_N(x) = \frac{1}{n\, r_k(x)^d} \sum_{i=1}^{n} \kappa\left(\frac{x - X_i}{r_k(x)}\right), \quad x \in \mathbb{R}^d, \tag{3.34}$$

   is called *generalized nearest neighbor estimator* of f.

2. The estimator $\hat{f}_V$ defined by

$$\hat{f}_V(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{r_k(X_i)^d}\, \kappa\left(\frac{x - X_i}{r_k(X_i)}\right), \quad x \in \mathbb{R}^d, \tag{3.35}$$

   is called *variable kernel estimator* of f.

Remark 3.4.5. Let $X_1, \ldots, X_n$ and $f : \mathbb{R}^d \to [0, \infty)$ be given as in Definition 3.4.3 and let $\hat{f}_N$ and $\hat{f}_V$ be the generalized nearest neighbor estimator and the variable kernel estimator introduced in (3.34) and (3.35).

1. The difference between $\hat{f}_N$ and $\hat{f}_V$ can be explained as follows. In the generalized nearest neighbor approach, the kernel functions that are assigned to the random vectors $X_1, \ldots, X_n$ are all scaled using the same smoothing parameter but this parameter changes when the PDF f is estimated at different locations in $\mathbb{R}^d$. In the variable kernel approach, the kernel functions that correspond to $X_1, \ldots, X_n$ are scaled using different bandwidths but this choice remains the same for all $x \in \mathbb{R}^d$.

2. As a consequence of the previous point, each realization of the variable kernel estimator $\hat{f}_V$ is a valid PDF, which, in general, is not the case for the generalized nearest neighbor estimator $\hat{f}_N$.

3. More general adaptive smoothing parameterizations are considered in literature by, e.g., using a different bandwidth matrix $H(x) \in \mathsf{H}$ for different locations $x \in \mathbb{R}^d$ (generalization of the nearest neighbor estimator) or by assigning a different bandwidth matrix $H_i \in \mathsf{H}$ to each random vector X_i for $i = 1, \ldots, n$ (generalization of the variable kernel estimator), see Scott (1992), Section 6.6 and Simonoff (1996), Section 4.3.1. However, this is rarely used in practice as it is almost impossible to properly choose all smoothing parameters.

3.4.3. Kernel regression

Another fundamental statistical problem that occurs in many fields of science and technology is the modeling and estimation of relationships between different variables. A popular method to address such kind of problems is regression analysis, which is, according to Scott (1992), the most commonly used statistical technique. Consider a sequence of $n \in \mathbb{N}$ independent and identically distributed absolutely continuous random vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $X_i = (X_{i1}, \ldots, X_{id})$ is a d-dimensional random vector and Y_i is a real-valued random variable for $i = 1, \ldots, n$. By $f : \mathbb{R}^{d+1} \to [0, \infty)$ we denote the PDF of the random vector (X_1, Y_1) and by $f_X : \mathbb{R}^d \to [0, \infty)$ the PDF of X_1. Regression analysis aims to describe how the random variable Y_i can be explained by the random vector X_i for $i = 1, \ldots, n$, which is why Y_i is denoted as *response variable* and $X_{i1}, \ldots, X_{id}$ are called *explanatory* or *regressor variables.* A general relationship is described by the *regression model*

$$Y_i = m(X_i) + v^{\frac{1}{2}}(X_i)\,\varepsilon_i, \quad i = 1, \ldots, n, \tag{3.36}$$

where the (unknown) *regression function* $m : \operatorname{supp}(f_X) \to \mathbb{R}$ given by

$$m(x) = \mathbb{E}\,(Y_1 \,|\, X_1 = x), \quad x \in \operatorname{supp}(f_X), \tag{3.37}$$

is the conditional expectation function of Y_1 given X_1 and $v : \operatorname{supp}(f_X) \to [0, \infty)$ with

$$v(x) = \operatorname{var}(Y_1 \,|\, X_1 = x), \quad x \in \operatorname{supp}(f_X), \tag{3.38}$$

is the conditional variance function of Y_1 given X_1. Furthermore, $\varepsilon_1, \ldots, \varepsilon_n$ are some independent and identically distributed random variables, called *residuals*, with $\mathbb{E}\,\varepsilon_i = 0$ and $\operatorname{var} \varepsilon_i = 1$ for $i = 1, \ldots, n$. Additionally, the residuals $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be independent of $X_1, \ldots, X_n$. Since the regressor variables are supposed to be random, the described setting is referred to as random design in literature. In contrast to this, a fixed design can be considered where the explanatory variables are deterministic but this is not discussed here.

A frequently considered approach (especially if $d = 1$) is to find a parametric representation of the regression function m and to estimate the regression parameters using least squares or maximum likelihood methods. In this context, the most popular choice is to assume m to be a linear function, which is referred to as linear regression, but a wide variety of other parametric models exists, see, e.g., Chatterjee and Hadi (2012). However, not in all applications the use of parametric models is possible (or reasonable), which is why several nonparametric regression methods have been developed, see, e.g., Takezawa (2006).

In the present thesis, we focus on kernel regression methods. Almost all (nonparametric) estimators $\hat{m}$ of the regression function m considered in literature are weighted (local) averages of the response variables $Y_1, \ldots, Y_n$, i.e.,

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x, X_1, \ldots, X_n)\, Y_i, \quad x \in \operatorname{supp}(f_X), \tag{3.39}$$

with nonnegative weights $w_1(x, X_1, \ldots, X_n), \ldots, w_n(x, X_1, \ldots, X_n) \geq 0$ such that $\sum_{i=1}^{n} w_i(x, X_1, \ldots, X_n) = 1$. A particularly intuitive choice of the weights can be derived as follows. For each $x \in \operatorname{supp}(f_X)$, let $f_{Y|X=x} : \mathbb{R} \to [0, \infty)$ be the conditional PDF of Y_1 given $\{X_1 = x\}$. Then, the conditional expectation function m can be represented as

$$m(x) = \int_{\mathbb{R}} y f_{Y|X=x}(y)\, \mathrm{d}y = \frac{1}{f_X(x)} \int_{\mathbb{R}} y f(x, y)\, \mathrm{d}y, \quad x \in \operatorname{supp}(f_X). \tag{3.40}$$

The only unknown quantities on the right-hand side of (3.40) are the PDFs f_X and f, which can be estimated using the kernel-based approaches introduced in Section 3.4.2. A KDE $\hat{f}_X$ of f_X is given by

$$\hat{f}_X(x) = \frac{1}{n\sqrt{\det(H)}} \sum_{i=1}^{n} \kappa\left(H^{-\frac{1}{2}}(x - X_i)\right), \quad x \in \mathbb{R}^d,$$

for some d-dimensional kernel function $\kappa : \mathbb{R}^d \to [0, \infty)$ and a bandwidth matrix $H \in \mathsf{H}$, compare to (3.32). A similar KDE $\hat{f}$ of f is given by

$$\hat{f}(x, y) = \frac{1}{n\sqrt{\det(H)}\, h} \sum_{i=1}^{n} \kappa\left(H^{-\frac{1}{2}}(x - X_i)\right) \tilde{\kappa}\left(\frac{y - Y_i}{h}\right), \quad x \in \mathbb{R}^d, y \in \mathbb{R},$$

where, additionally to the notation introduced above, $\tilde{\kappa} : \mathbb{R} \to [0, \infty)$ denotes a one-dimensional kernel function and $h > 0$ is the corresponding bandwidth. Using the properties of a kernel function stated in (3.27) and (3.28) we get that

$$\begin{aligned}
\int_{\mathbb{R}} y \hat{f}(x, y)\, \mathrm{d}y &= \frac{1}{n\sqrt{\det(H)}} \sum_{i=1}^{n} \kappa\left(H^{-\frac{1}{2}}(x - X_i)\right) \int_{\mathbb{R}} \frac{y}{h} \tilde{\kappa}\left(\frac{y - Y_i}{h}\right) \mathrm{d}y \\
&= \frac{1}{n\sqrt{\det(H)}} \sum_{i=1}^{n} \kappa\left(H^{-\frac{1}{2}}(x - X_i)\right) \int_{\mathbb{R}} (Y_i + hz) \tilde{\kappa}(z)\, \mathrm{d}z \\
&= \frac{1}{n\sqrt{\det(H)}} \sum_{i=1}^{n} Y_i\, \kappa\left(H^{-\frac{1}{2}}(x - X_i)\right), \quad x \in \mathbb{R}^d.
\end{aligned}$$

Consequently, replacing f_X and f by the KDEs $\hat{f}_X$ and $\hat{f}$ in (3.40) leads to the following estimator of the regression function m, which has first been introduced in Nadaraya (1964) and Watson (1964).

Definition 3.4.4 (Nadaraya-Watson estimator)**.** Consider the regression model specified in (3.36) with regression function $m : \mathrm{supp}(f_X) \to \mathbb{R}$ as given in (3.37). Furthermore, let $H \in \mathsf{H}$ be a (symmetric and positive definite) bandwidth matrix and $\kappa : \mathbb{R}^d \to [0, \infty)$ a d-dimensional kernel function. The estimator $\hat{m}_{NW}$ defined by

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} Y_i \, \kappa\left(H^{-\frac{1}{2}}(x - X_i)\right)}{\sum_{i=1}^{n} \kappa\left(H^{-\frac{1}{2}}(x - X_i)\right)}, \quad x \in \mathrm{supp}(f_X), \tag{3.41}$$

is called *Nadaraya-Watson estimator* (NWE) of m. If κ has bounded support, the denominator in (3.41) can be equal to zero, which implies that the numerator is equal to zero as well. In this case, the estimator $\hat{m}_{NW}(x)$ is not defined.

Remark 3.4.6. Let $m : \mathrm{supp}(f_X) \to \mathbb{R}$ be given as in Definition 3.4.4 and let $\hat{m}_{NW}$ be the NWE introduced in (3.41).

1. In fact, the estimator $\hat{m}_{NW}(x)$ of m at $x \in \mathrm{supp}(f_X)$ is a weighted average of the response variables $Y_1, \dots, Y_n$ as specified in (3.39) with nonnegative weights $w_1(x, X_1, \dots, X_n), \dots, w_n(x, X_1, \dots, X_n)$ given by
$$w_j(x, X_1, \dots, X_n) = \frac{\kappa\left(H^{-\frac{1}{2}}(x - X_j)\right)}{\sum_{i=1}^{n} \kappa\left(H^{-\frac{1}{2}}(x - X_i)\right)}, \quad j = 1, \dots n.$$

2. The NWE is a special case of a much broader class of estimators called local polynomial kernel estimators, see, e.g., Wand and Jones (1995), Section 5.2. In this context, the NWE is denoted as local polynomial kernel estimator of degree 0 or as locally constant kernel estimator.

3. The bandwidth matrix $H \in \mathsf{H}$ controls the influence the response variables $Y_1, \dots, Y_n$ have on the estimator $\hat{m}_{NW}(x)$ for $x \in \mathrm{supp}(f_X)$. If, for example, the kernel function $\kappa : \mathbb{R}^d \to [0, \infty)$ is radially symmetric with support $b(o, 1)$ and the bandwidth matrix H is given as $H = h^2 I$ for some $h > 0$, then $\hat{m}_{NW}(x)$ only takes into account those response variables that fall into $b(x, h)$.

4. Let the kernel function κ be defined by $\kappa(x) = c_d^{-1} \kappa_1(\sqrt{x^\top x})$ for $x \in \mathbb{R}^d$ with a normalizing constant $c_d > 0$ and a one-dimensional kernel $\kappa_1 : \mathbb{R} \to [0, \infty)$ and let the bandwidth matrix H be given as $H = h^2 I$ for some $h > 0$. Then, the NWE $\hat{m}_{NW}$ involves the computation of d-dimensional Euclidean distances, compare to Remark 3.4.3. When considering a geographical domain $\mathcal{T}$ instead of $\mathbb{R}^d$, then it is more suitable to use the great-circle distance introduced in (3.4) rather than the d-dimensional Euclidean distance. However, an appropriate adjustment of the normalizing constant c_d is not necessary since c_d is canceled out in the definition of the NWE.

To conclude the discussion of the NWE, we briefly examine some statistical properties. Since a random design is considered, i.e., the regressors $X_1, \ldots, X_n$ are random vectors, there often is a positive probability of the estimator $\hat{m}_{NW}(x)$ being undefined due to the denominator in (3.41) being equal to 0 if a kernel function with bounded support is used. In this case, the moments of $\hat{m}_{NW}(x)$ are undefined, too. For that reason, Ruppert and Wand (1994) suggest to derive approximate formulas for the conditional expectation and variance of $\hat{m}_{NW}(x)$ given $\{X_1, \ldots, X_n\}$ for certain $x \in \operatorname{supp}(f_X)$. For simplicity, we consider the bandwidth matrix of the NWE to be a multiple of the d-dimensional unit matrix. The more general case is discussed in Härdle et al. (2004), Section 4.5.

Theorem 3.4.2. *We consider the regression model specified in (3.36) with regression function $m : \operatorname{supp}(f_X) \to \mathbb{R}$ as given in (3.37). Suppose that the PDF $f_X : \mathbb{R}^d \to [0, \infty)$ of the regressors $X_1, \ldots, X_n$ and the regression function m have continuous first- and second-order partial derivatives and that the conditional variance function $v : \operatorname{supp}(f_X) \to [0, \infty)$ specified in (3.38) is continuous. Furthermore, let $\hat{m}_{NW}$ be the NWE of m with bandwidth matrix $H = h^2 I$ for some $h > 0$ and kernel function $\kappa : \mathbb{R}^d \to [0, \infty)$, which is assumed to be a radially symmetric kernel or a product kernel with compact support. Finally, let x be in the interior of $\operatorname{supp}(f_X)$ with $v(x) > 0$. If n is large, h is small and nh^d is large, then*

(i) the conditional expectation of $\hat{m}_{NW}(x)$ given $\{X_1, \ldots, X_n\}$ can be approximated by

$$\mathbb{E}\left(\hat{m}_{NW}(x) \,|\, X_1, \ldots, X_n\right) \approx m(x) + h^2 \mu_2(\kappa) \left(\frac{\nabla_m(x)^\top \nabla_{f_X}(x)}{f_X(x)} + \frac{\operatorname{tr}(\mathcal{H}_m(x))}{2} \right),$$

(ii) the conditional variance of $\hat{m}_{NW}(x)$ given $\{X_1, \ldots, X_n\}$ can be approximated by

$$var\left(\hat{m}_{NW}(x) \,|\, X_1, \ldots, X_n\right) \approx \frac{1}{nh^d} \frac{v(x)}{f_X(x)} \int_{\mathbb{R}^d} \kappa^2(y)\, \mathrm{d}y,$$

where $\nabla_m(x)$ and $\nabla_{f_X}(x)$ are the gradients of m and f_X at x and $\mathcal{H}_m(x)$ is the Hessian matrix of m at x introduced in Section 3.1.1.

Proof. For *(i)* see Härdle et al. (2004), Section 4.5.1 and for *(ii)* see Ruppert and Wand (1994), Section 2.

3.4.4. Selection of smoothing parameters

One of the most challenging tasks in both kernel density estimation and kernel regression is to suitably choose the smoothing parameters. In many applications it is reasonable to find an optimal bandwidth matrix $H \in \mathsf{H}$ by comparing estimates for different choices of H and selecting the matrix that provides the best compromise between eliminating noise and showing enough details of the estimated function. In some cases it is even possible to find a physical meaning of the bandwidth matrix H in the considered application according to which H can be selected. However, in a majority of situations this is not possible and one aims to algorithmically determine the smoothing parameters based on the underlying data. In the following, we present some basic approaches that are considered in this thesis, where we start with the selection of a bandwidth matrix for a KDE.

Let $X_1, \ldots, X_n$ with $n \in \mathbb{N}$ be a sequence of independent and identically distributed absolutely continuous random vectors in $\mathbb{R}^d$ with PDF $f : \mathbb{R}^d \to [0, \infty)$ and let $\hat{f}_K$ be the KDE introduced in (3.32) with bandwidth matrix $H \in \mathsf{H}$ and kernel function $\kappa : \mathbb{R}^d \to [0, \infty)$. A first approach is based on the consideration that a reasonable choice of H should lead to a small, if possible minimal, approximate MISE of $\hat{f}_K$. However, the bandwidth matrix minimizing the approximate MISE can only be derived under certain constraints. We assume f to be the PDF of the multivariate normal distribution with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ for $\sigma_1^2, \ldots, \sigma_d^2 > 0$ and κ to be the d-dimensional Gaussian kernel. Then, it can be shown that among all diagonal bandwidth matrices $\{H = \operatorname{diag}(h_1^2, \ldots, h_d^2) : h_1^2, \ldots, h_d^2 > 0\}$ the one minimizing the approximate MISE of $\hat{f}_K$ is given by

$$h_j = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{\frac{-1}{d+4}} \sigma_j, \quad j = 1, \ldots, d,$$

see, e.g., Härdle et al. (2004), Section 3.6.2. Accordingly, plug-in estimators $\hat{h}_1, \ldots, \hat{h}_d$ of the optimal bandwidths $h_1, \ldots, h_d$ are given by

$$\hat{h}_j = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{\frac{-1}{d+4}} \hat{\sigma}_j, \quad j = 1, \ldots, d, \tag{3.42}$$

where $\hat{\sigma}_1, \ldots, \hat{\sigma}_d$ are standard moment estimators of $\sigma_1, \ldots, \sigma_d$. Choosing the bandwidths according to (3.42) is denoted as *normal reference rule* in literature. Since the term $\left(\frac{4}{d+2}\right)^{\frac{1}{d+4}}$ is approximately equal to 1 for all $d \in \mathbb{N}$ (exactly equal to 1 if $d = 2$), this term is often dropped. The resulting simplified version of (3.42) is referred to as *Scott's rule*, see, e.g., Scott (1992), Section 6.3.1. Note that the case of f being the PDF of a normal distribution is not interesting in applications as in that situation, f would be estimated by determining its parameters using the method of moments. However, Härdle et al. (2004) argue that the normal reference rule and Scott's rule still provide bandwidths that are not too far from being optimal if f is unimodal, fairly symmetric and has non-heavy tails. In all other situations the rules are still applicable but tend to provide slightly too large bandwidths causing moderate oversmoothing.

A popular but costly alternative to the simple rules suggested above is the application of likelihood cross-validation (LCV), see, e.g., Silverman (1986), Section 3.4.4. Further cross-validation methods such as least-squares cross-validation or biased cross-validation exist, see, e.g., Härdle (1991), but are not considered in this thesis. The idea behind LCV is the following. Suppose that in addition to $X_1, \ldots, X_n$ a sample variable X (independent of $X_1, \ldots, X_n$) with PDF f is given. Then, it is desirable to choose the bandwidth matrix $H \in \mathsf{H}$ such that the likelihood $\log(\hat{f}_K(X))$ (as a function of H) is large. However, an additional sample variable X is not available, which is why it is suggested to omit one random variable X_i of the original sample, to construct an estimator $\hat{f}_K^{(-i)}$ based on $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ according to (3.32), and to consider the likelihood $\log(\hat{f}_K^{(-i)}(X_i))$ (still depending on $H \in \mathsf{H}$). This is done for all $i = 1, \ldots, n$ and the (random) likelihood function $L : \mathsf{H} \to \mathbb{R}$ is defined as

$$L(H) = \sum_{i=1}^{n} \log\left(\hat{f}_K^{(-i)}(X_i)\right), \quad H \in \mathsf{H},$$

where the estimators $\hat{f}_K^{(-i)}$ technically are functions on H for $i = 1, \ldots, n$. Finally, an estimator $\hat{H}$ of an optimal bandwidth matrix H is given by

$$\hat{H} = \operatorname*{argmax}_{H \in \mathsf{H}} L(H).$$

However, LCV is very sensitive to outliers and, as mentioned before, computationally expensive if H has a high number of independent entries or if the number n of sample variables is large.

A similar cross-validation technique can be used to select a bandwidth matrix in kernel regression (although it is not denoted as LCV in this context). The one-dimensional case is discussed in Härdle et al. (2004), Section 4.3.2 but a generalization to multivariate kernel regression is straightforward. Consider the regression model specified in (3.36) with regression function as given in (3.37). The general proceeding of the cross-validation method is similar to LCV. For each $i \in \{1, \ldots, n\}$, let $\hat{m}_{NW}^{(-i)}$ be a NWE defined according to (3.41) based on the regressor variables $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ and the response variables $Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n$. As it is desired that the estimator $\hat{m}_{NW}^{(-i)}(X_i)$ (as a function of $H \in \mathsf{H}$) is close to the response variable Y_i for all $i = 1, \ldots, n$, we consider the (random) cross-validation function $CV : \mathsf{H} \to [0, \infty)$ given by

$$CV(H) = \sum_{i=1}^{n} \left(Y_i - \hat{m}_{NW}^{(-i)}(X_i)\right)^2$$

and define an estimator $\hat{H}$ of the optimal bandwidth matrix H as

$$\hat{H} = \operatorname*{argmin}_{H \in \mathsf{H}} CV(H).$$

Similar to LCV in density estimation, this cross-validation method can be computationally expensive for high-dimensional bandwidth matrices or large sample sizes. Further approaches to the selection of an optimal bandwidth matrix in kernel regression exist but are not applied in this thesis, see, e.g., Härdle (1991), Chapter 6 for more details.

Part II.

Stochastic models in probabilistic weather prediction

4. Stochastic model for the occurrence of precipitation

In the present chapter, we introduce a probabilistic approach to the computation of area probabilities for the occurrence of precipitation based on a spatial stochastic model of precipitation cells. As mentioned in Section 1.3, this method was developed in the context of a research collaboration with DWD. The proposed model of precipitation is of fundamental importance for the rest of Part II of this thesis as it constitutes a basis for subsequent modeling approaches on precipitation amounts and thunderstorm cells in Chapters 5 and 6. The underlying data provided by DWD are described in Section 4.1. A suitable mathematical framework for the modeling of point probabilities included in the data is introduced in Sections 4.2 and 4.3. In Section 4.4, a stochastic model of precipitation cells is developed, whereas statistical methods for the computation of model characteristics are presented in Section 4.5. In Section 4.6, it is demonstrated how area probabilities of precipitation can be computed or estimated according to the specified model of precipitation cells. Furthermore, examples of simulated precipitation patterns and estimated area probabilities are illustrated. To conclude, Section 4.7 provides a comparison of point and area probabilities with radar-derived precipitation analyses using typical score functions and diagrams from weather forecast verification. The results presented in this chapter (in a slightly modified version) have been incorporated in Kriesche et al. (2015a). The mathematical framework as introduced in Sections 4.2 and 4.3 has been described in Kriesche et al. (2017b) and Kriesche et al. (2017c). Although we focus on the weather event 'occurrence of precipitation' in this chapter, the proposed methodology can generally be used to compute area probabilities for the occurrence of other weather events, too.

4.1. Description of data

The model-based approach to the computation of area probabilities presented in this chapter relies on point probabilities for the occurrence of precipitation as its sole data input. We consider a system of 503 German and Luxembourgian synoptic weather stations, which are located in the window $[5.3°\mathrm{E}, 15.3°\mathrm{E}] \times [46.9°\mathrm{N}, 55.3°\mathrm{N}]$. Technically, this window describes a geographical domain and distances between arbitrary locations

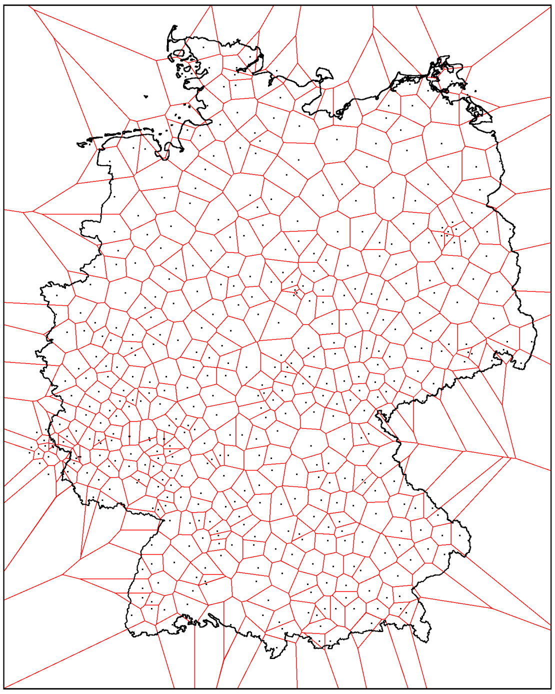
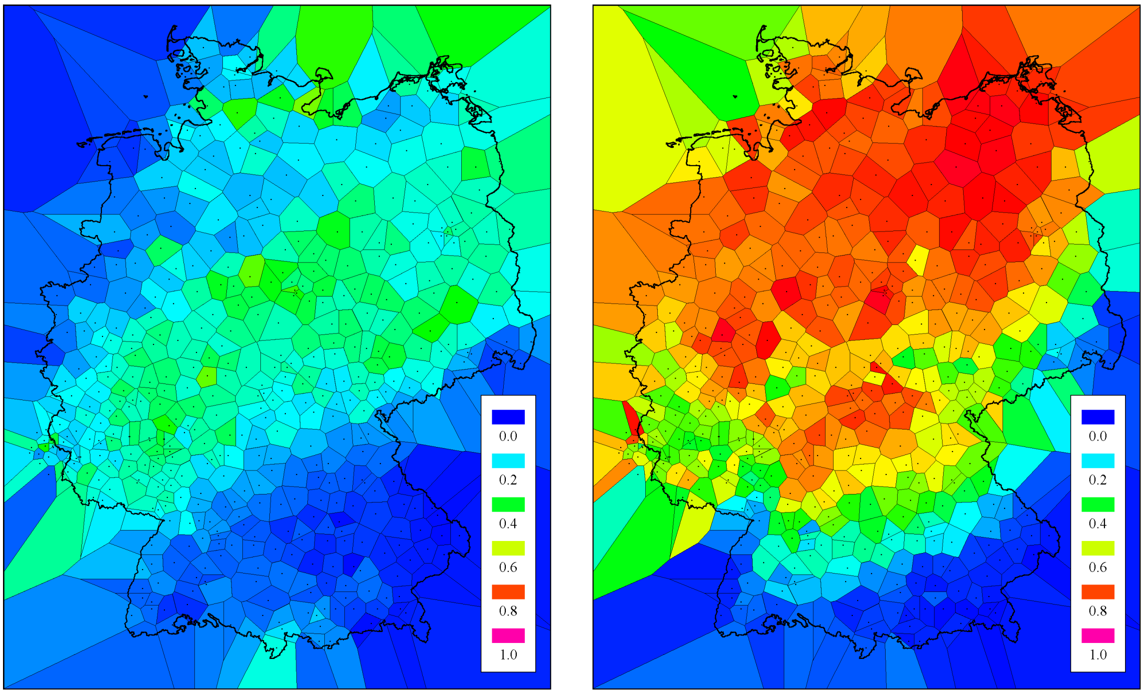


Figure 4.1.: Locations of 503 weather stations at which point probabilities of precipitation are available together with the corresponding Voronoi tessellation.

need to be computed using the great-circle distance introduced in (3.4). This, however, would make the application of the spatial models and methods that are used in this chapter considerably more complicated or even impossible. Thus, we project the window $[5.3°\mathrm{E}, 15.3°\mathrm{E}] \times [46.9°\mathrm{N}, 55.3°\mathrm{N}]$ including the locations of the 503 weather stations to the rectangle $[0, 1500] \times [0, 1875]$ and perform all computations using the two-dimensional Euclidean distance in the following. By doing so we introduce an error due to ignoring the curvature of the earth's surface but this has a negligible effect for such a relatively small area as the one considered here. Locations in the rectangle $[0, 1500] \times [0, 1875]$ and distances between them are denoted as pixel coordinates and pixel distances (or simply pixels) in the following. However, the rectangle is chosen in such a way that any pixel distance $d > 0$ approximately represents a distance of $0.5 \cdot d$ km. In Figure 4.1, the described rectangle including the 503 weather stations and the corresponding Voronoi tessellation, see (4.2) later on, are shown.

The data cover a time frame of four months in the year 2012. In order to address seasonal changes in precipitation patterns, a summer period from June 1 until July 31 and a winter period from November 1 until December 31 were selected. At each day, forecasts were made at 12 am (midnight) for seven forecast periods of one hour each, ranging from 2-3 UTC every three hours up to 20-21 UTC, which results in a total of 854 available forecast periods. For each such period, re-forecasts of the

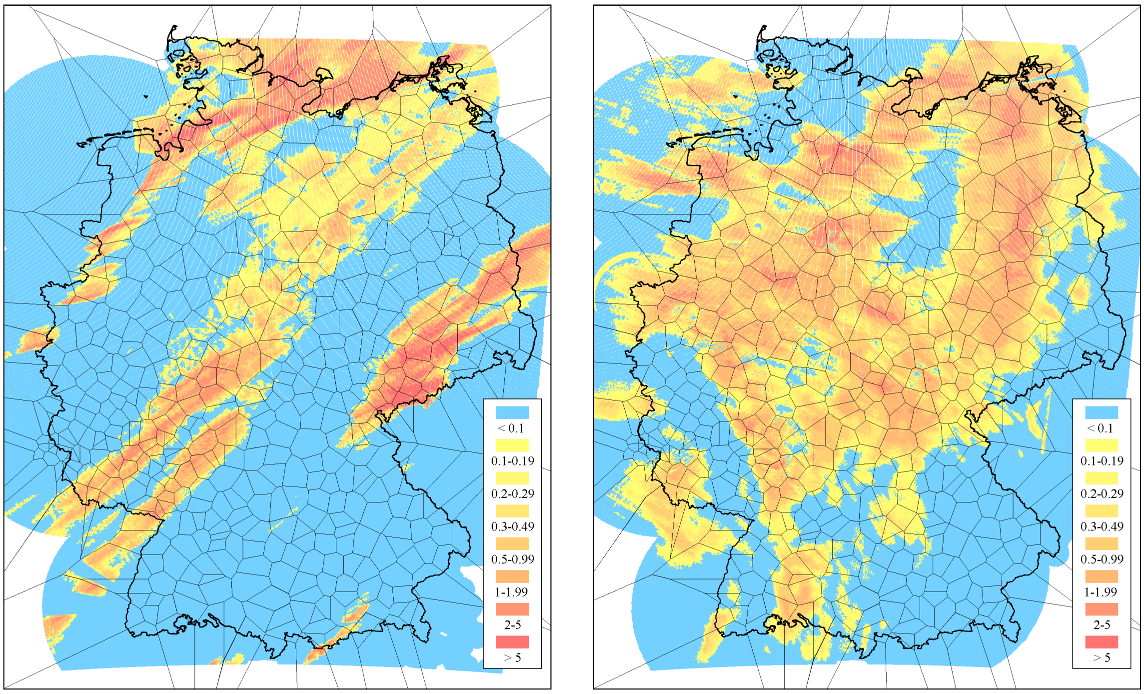


June 16, 2012, 20-21 UTC   December 9, 2012, 11-12 UTC

Figure 4.2.: Available point forecast data for two selected forecast periods: the locations of the considered 503 weather stations and the corresponding Voronoi tessellation, where each Voronoi cell is colored according to the point probability of precipitation at the corresponding weather station.

MOSMIX system of DWD were started, where a mixture of the individually post-processed forecasts of the GME and the IFS of ECMWF were used, see Section 2.3.2, to provide point probabilities of precipitation exceeding an amount of 0.1 mm for all 503 weather stations. The threshold of 0.1 mm is chosen in order to allow for a consistent verification with rain gauge adjusted radar analyses, which are not able to detect smaller precipitation amounts. Figure 4.2 illustrates point probabilities for two sample forecast periods: June 16, 2012, 20-21 UTC, where moderate probabilities are forecasted for a band ranging from the west of Germany to the northeast, and December 9, 2012, 11-12 UTC, with high probabilities in northern and central Germany, which gradually decrease towards the south.

For the purpose of forecast verification radar-derived precipitation analyses from RADOLAN (see Section 2.1) are available for all forecast periods considered above. RADOLAN observations cover the entire territory of Germany except the northern part of the island of Sylt, where one weather station is located. Accordingly, forecast verification should only be performed for points and areas inside the boundaries of



June 16, 2012, 20-21 UTC December 9, 2012, 11-12 UTC

Figure 4.3.: Adjusted precipitation amounts in mm from RADOLAN for two selected forecast periods.

mainland Germany. Adjusted precipitation amounts from RADOLAN have a limited resolution as only values being a multiple of 0.1 mm are measurable. Therefore, in this chapter the occurrence of precipitation at some location is identical with a precipitation amount of more than 0.1 mm. Adjusted radar data for two sample forecast periods are illustrated in Figure 4.3. A comparison with point probabilities in Figure 4.2 reveals that in the considered periods precipitation mainly occurred in those regions with higher point probabilities indicating a close correspondence of both kind of data.

4.2. Underlying probability space

Before describing an approach to the spatial stochastic modeling of precipitation patterns with the purpose of estimating area probabilities for the occurrence of precipitation, we need to specify a suitable mathematical framework. In the following, we always consider a fixed one-hour forecast period T that is interpreted as a subinterval of the real line with a length of 60 minutes. In particular, no temporal dynamic is taken into account in the proposed modeling approach. The occurrence of precipitation is one of the most complex meteorological events, which makes a precise estimation of probabilities extremely difficult. Point forecasts computed based on NWP models and ensemble prediction systems are subject to several sources of uncertainty concerning, e.g., initial weather conditions or inaccuracies in the model specification due to discretization and physical parameterization. Thus, these estimated point probabilities are subject to both random and systematic errors, see Sections 2.2 and 2.3.1. In order to eliminate systematic errors, post-processing methods such as MOS are applied by DWD, which provide statistically unbiased probabilistic forecasts, see Section 2.3.2. These forecasts, however, are still subject to random errors and can thus be interpreted as estimators of the unknown future weather conditions. To provide a suitable framework, we introduce the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is an abstract space describing all possible weather scenarios and the corresponding forecasts provided by the (numerical and probabilistic) weather forecast models of DWD, $\mathcal{F}$ denotes a suitable σ-algebra of subsets of Ω and $\mathbb{P}$ is some probability measure on $(\Omega, \mathcal{F})$, which associates each event $A \in \mathcal{F}$ with the probability $\mathbb{P}(A) \in [0,1]$ of its occurrence. As mentioned previously, probabilistic forecasts are subject to a random error, which is modeled by a random variable $E : \Omega \to \mathbb{S}$, with $\mathbb{S}$ being the (abstract) measurable set of all such possible errors. Accordingly, the error E can also be interpreted as the deviation of the forecast from the actual future weather situation. Heuristically spoken, conditioning on $\{E = e\}$ for some realization $e \in \mathbb{S}$ of E is synonymous for having a specific forecast provided by the models of DWD (with error e) available. This is always the case in applications, which is why model fitting procedures and simulation algorithms are typically described conditioned on $\{E = e\}$ in this thesis.

4.3. Modeling of point probabilities

We introduce our model-based method for the computation of area precipitation probabilities in a general context (i.e., not restricted to the data described in Section 4.1) in order to allow for a flexible application. As explained in Sections 2.3 and 4.2, reliable, unbiased point forecasts (including point probabilities for the occurrence of precipitation) are derived by DWD using numerical models, ensemble prediction systems, and statistical post-processing. Since these forecasts are interpreted as estimators, we model point probabilities as random variables. For that purpose, consider a random field $\{P(t), t \in W\}$ in a compact and convex domain $W \subset \mathbb{R}^2$, where $P(t) : \Omega \to [0,1]$ denotes the random point probability for the occurrence of precipitation at location $t \in W$ within the considered forecast period T. For each $t \in W$, we assume the random variable $P(t)$ to be $\sigma(E)$-measurable, where $\sigma(E) \subset \mathcal{F}$ is the sub-σ-algebra of events generated by the random error E. This is equivalent to the existence of a measurable function $f : \mathbb{S} \to [0,1]$ such that $P(t) = f(E)$, see, e.g., Jacod and Protter (2004), Theorem 23.2. Accordingly, if conditioned on $\{E = e\}$ for any realization e of E, the value of $P(t)$ is non-random and only depends on e. In this case, any realization of $P(t)$ can be identified by the conditional expectation $f(e) = \mathbb{E}\,(P(t)\,|\,E = e)$. In general, point probabilities can be estimated by DWD for any location t inside the domain W. In practice, however, this is only done for a finite number of sites (e.g., a network of weather stations or a regularly spaced lattice). Accordingly, we suppose that for some $n \in \mathbb{N}$ there is a finite sequence $s_1, \ldots, s_n \in W$ of n geographically distinct locations at which point probabilities $p(s_1) = \mathbb{E}\,(P(s_1)\,|\,E = e), \ldots, p(s_n) = \mathbb{E}\,(P(s_n)\,|\,E = e)$ are available for a particular realization e of E (i.e., from a specific forecast provided by the models of DWD). In the example of application discussed in this thesis to illustrate the proposed methodology, we have that $W = [0, 1500] \times [0, 1875]$ denotes a rectangular domain comprising the boundaries of Germany and Luxembourg, the sequence $s_1, \ldots, s_{503}$ identifies the locations of the $n = 503$ weather stations shown in Figure 4.1 and $p(s_1), \ldots, p(s_{503})$ denotes the available point probabilities for the occurrence of precipitation described in Section 4.1, where e is the error that occurs when computing these data.

A fundamental assumption in the modeling of probabilities for the occurrence of precipitation is that a precipitation pattern consist of several precipitation cells and that there is precipitation at some location $t \in W$ if and only if t is covered by at least one such cell. We model the union set of precipitation cells using a random closed set $M : \Omega \to \mathsf{C}$, see Section 3.3.6. Consequently, the random field $\{P(t), t \in W\}$ of point probabilities is represented as

$$P(t) = \mathbb{P}(t \in M \,|\, E), \quad t \in W. \tag{4.1}$$

In order to give a more precise representation of point probabilities and further probabilistic characteristics such as, e.g., area probabilities, a model for the random closed set M of precipitation cells needs to be found.

4.4. Spatial stochastic model for precipitation cells

In this section, the random closed set M of precipitation cells is further specified. We start with the modeling of cell centers. One major requirement for a stochastic model to be applied in operational weather prediction is spatial non-stationarity to account for geographical differences (e.g., plains in northern Germany, uplands in central Germany, mountains in the south) as well as locally varying weather conditions in the considered forecast period T. For this purpose, we suppose that precipitation cells, or more precisely their cell centers, occur in W according to a random intensity function $\{\Lambda(t), t \in W\}$, where $\Lambda(t) : \Omega \to [0, \infty)$ is a nonnegative random variable modeling the intensity for the occurrence of a precipitation cell at location $t \in W$. Analogous to the random point probabilities $\{P(t), t \in W\}$, the intensity $\Lambda(t)$ is assumed to be

$\sigma(E)$-measurable for each $t \in W$, i.e., $\Lambda(t)$ is non-random conditioned on $\{E = e\}$ for any $e \in \mathbb{S}$. It turns out to be complicated to find a model of $\{\Lambda(t), t \in W\}$ with smooth realizations that captures the degree of non-stationarity sufficiently well but can still be fitted in a reasonable way based on point forecast data. To account for the circumstance that point probabilities are only available at the sites $s_1, \ldots, s_n \in W$, we make the simplifying assumption that realizations of $\{\Lambda(t), t \in W\}$ are piecewise constant in a neighborhood of each site s_i for $i = 1, \ldots, n$. The most natural choice of such a neighborhood is the Voronoi tessellation $\{V(s_1), \ldots, V(s_n)\}$ of $s_1, \ldots, s_n$ in W, where the Voronoi cell $V(s_i)$ of s_i is defined as

$$V(s_i) = \{x \in W : \|x - s_i\|_2 < \|x - s_j\|_2 \text{ for all } j = 1, \ldots, n \text{ with } j \neq i\}, \tag{4.2}$$

for $i = 1, \ldots, n$. Consequently, we assume that the random intensity function $\{\Lambda(t), t \in W\}$ for the occurrence of precipitation cells can be represented as

$$\Lambda(t) = \sum_{j=1}^{n} A_j \, \mathbb{1}_{V(s_j)}(t), \quad t \in \bigcup_{i=1}^{n} V(s_i),$$

where the $\sigma(E)$-measurable random variables $A_1, \ldots, A_n : \Omega \to [0, \infty)$ can be interpreted as random local intensities for the occurrence of precipitation cells in neighborhoods of $s_1, \ldots, s_n$. If $t \in W$ is located on the boundary of at least one Voronoi cell, we set $\Lambda(t)$ equal to the minimum intensity of all adjacent Voronoi cells. Having specified the intensity function $\{\Lambda(t), t \in W\}$, a model for the centers of precipitation cells can be given. We suggest to use a two-dimensional Cox point process $\{X_i, i = 1, \ldots, Z\}$ for that purpose (see Section 3.3.4), where $Z : \Omega \to \{0, 1, \ldots\}$ is a random variable describing the total number of precipitation cells in W. In particular, the random variable Z is almost surely finite. Clearly, the Cox process $\{X_i, i = 1, \ldots, Z\}$ of cell centers cannot be assumed to be $\sigma(E)$-measurable because given a specific forecast of the weather forecast models of DWD, the weather activity in the (future) forecast period T (including the occurrence of precipitation) is still considered to be random.

After modeling cell centers, the shape of precipitation cells needs to be addressed. Due to the irregularity and complexity of precipitation fields (see Figure 4.3), it seems to be hardly possible to find a model for the shape of precipitation cells, which exactly matches real precipitation fields and, simultaneously, is still easy to handle concerning, e.g., model fitting and simulation. Therefore, we make the simplifying model assumption that there is precipitation at a location $t \in W$ (i.e., t is covered by at least one precipitation cell, see Section 4.3) if t has a distance of not more than R to at least one precipitation cell center, where $R : \Omega \to (0, \infty)$ is a $\sigma(E)$-measurable random variable denoted as precipitation range. Equivalently, the random closed set M of precipitation cells is represented as a germ-grain model based on $\{X_i, i = 1, \ldots, Z\}$

with circular grains and grain radius R, i.e.,

$$M = \bigcup_{i=1}^{Z} b(X_i, R). \tag{4.3}$$

Although it is obvious that precipitation cells are typically not circular in real precipitation patterns, they are often approximated as circular or elliptical discs in the literature, see Section 2.5. Thus, the grains $b(X_1, R), \dots, b(X_Z, R)$ are interpreted as models of precipitation cells and we denote the precipitation range R as random cell radius in this context. We want to emphasize that precipitation probabilities which are obtained as coverage probabilities of the germ-grain model M can still be quite accurate even if single realizations of M look atypical compared to observed precipitation fields from radar data. Analogous to the Cox process $\{X_i, i = 1, \dots, Z\}$ of cell centers, the germ-grain model M cannot be assumed to be $\sigma(E)$-measurable. Instead, conditioned on $\{E = e\}$ for any realization e of E, the Cox process $\{X_i, i = 1, \dots, Z\}$ is a Poisson process with (deterministic) intensity function $\{\lambda(t), t \in W\}$, see Section 3.3.3, where $\lambda(t) = \mathbb{E}\,(\Lambda(t) \,|\, E = e)$ for $t \in W$. Accordingly, the germ-grain model M is a Boolean model based on the Poisson process $\{X_i, i = 1, \dots, Z\}$ with circular grains having grain radius $r = \mathbb{E}\,(R \,|\, E = e)$ in this case, compare to Example 3.3.2.

Based on the characterization of M as a germ-grain model in (4.3), a more specific representation formula of the random field $\{P(t), t \in W\}$ of point probabilities for the occurrence of precipitation can be derived. Using the distributional properties of the Poisson process stated in (3.23) and the representation of $P(t)$ as a (conditional) coverage probability of M in (4.1) implies that

$$\begin{aligned}
P(t) &= \mathbb{P}(t \in M \,|\, E) \\
&= 1 - \mathbb{P}\big(\#\{i : X_i \in b(t, R)\} = 0 \,|\, E\big) \\
&= 1 - \exp\left\{ -\int_{b(t,R)} \sum_{i=1}^{n} A_i \, \mathbb{1}_{V(s_i)}(z) \,\mathrm{d}z \right\} \\
&= 1 - \exp\left\{ -\sum_{i=1}^{n} A_i \int_{W} \mathbb{1}_{b(t,R) \cap V(s_i)}(z) \,\mathrm{d}z \right\} \\
&= 1 - \exp\left\{ -\sum_{i=1}^{n} A_i \, \nu_2\big(b(t, R) \cap V(s_i)\big) \right\}, \quad t \in W.
\end{aligned} \tag{4.4}$$

4.5. Computation of model characteristics

Using the proposed model for the union of precipitation cells we can either compute area probabilities for the occurrence of precipitation directly or estimate these probabilities (and further characteristics) based on repeated Monte Carlo simulation, see Section 4.6. However, in order to do this, the random model characteristics $A_1, \ldots, A_n$ and R need to be determined first. As the proposed method is designed to be applicable in operational weather prediction, it is desirable that all characteristics are derived algorithmically from the random point probabilities $P(s_1), \ldots, P(s_n)$ without further input of the forecaster. In applications, the computation of model characteristics is performed in dependence of the available point probabilities $p(s_1), \ldots, p(s_n)$, which are computed based on a particular realization e of the random error E that results from the weather forecast models of DWD (i.e., $p(s_1) = \mathbb{E}\,(P(s_1)\,|\,E = e), \ldots, p(s_n) = \mathbb{E}\,(P(s_n)\,|\,E = e)$). By $a_1 = \mathbb{E}\,(A_1\,|\,E = e), \ldots, a_n = \mathbb{E}\,(A_n\,|\,E = e)$ and $r = \mathbb{E}\,(R\,|\,E = e)$ we denote the corresponding realizations of the random local intensities $A_1, \ldots, A_n$ for the occurrence of precipitation cells and the random cell radius R.

A simultaneous computation of the unknown characteristics $a_1, \ldots, a_n$ and r does not seem to be possible. Therefore, this section describes a multi-step procedure that can be outlined as follows.

1. For each $r' > 0$, intensities $a_1^{(r')}, \ldots, a_n^{(r')}$ are determined based on $p(s_1), \ldots, p(s_n)$ under the condition that precipitation cells have radius r'.
2. The cell radius r is computed as a function of $p(s_1), \ldots, p(s_n)$ and the family of conditional intensities $\{a_1^{(r')}, \ldots, a_n^{(r')}, r' > 0\}$.
3. Finally, the intensities $a_1, \ldots, a_n$ are obtained as $a_1^{(r)}, \ldots, a_n^{(r)}$ by setting $r' = r$ in step 1.

The single steps are described in detail in the following subsections.

4.5.1. Local intensities for the occurrence of precipitation cells

At first, we describe an approach to the computation of local intensities for the occurrence of precipitation cells with arbitrary radius based on the available point probabilities $p(s_1), \ldots, p(s_n)$. To be more precise, we aim to find a model configuration that provides point probabilities being as close as possible to the available data. For that purpose, it is assumed that for each $r' > 0$ there is a sequence of nonnegative intensities $a_1^{(r')}, \ldots, a_n^{(r')} \geq 0$, such that

$$p(s_j) = 1 - \exp\left\{ -\sum_{i=1}^{n} a_i^{(r')}\, \nu_2\big(b(s_j, r') \cap V(s_i)\big) \right\}, \quad j = 1, \ldots, n, \tag{4.5}$$

compare to (4.4). However, note that the characteristics $a_1^{(r')}, \ldots, a_n^{(r')}$ and r' are not necessarily suitable to describe point probabilities for locations $t \notin \{s_1, \ldots, s_n\}$, as this is only assumed to be the case if $a_i^{(r')} = a_i$ for $i = 1, \dots, n$ and $r' = r$. The system of equations stated in (4.5) can easily be transformed into

$$\log\left(\frac{1}{1-p(s_j)}\right) = \sum_{i=1}^{n} a_i^{(r')}\, \nu_2\big(b(s_j, r') \cap V(s_i)\big), \quad j = 1, \dots, n, \tag{4.6}$$

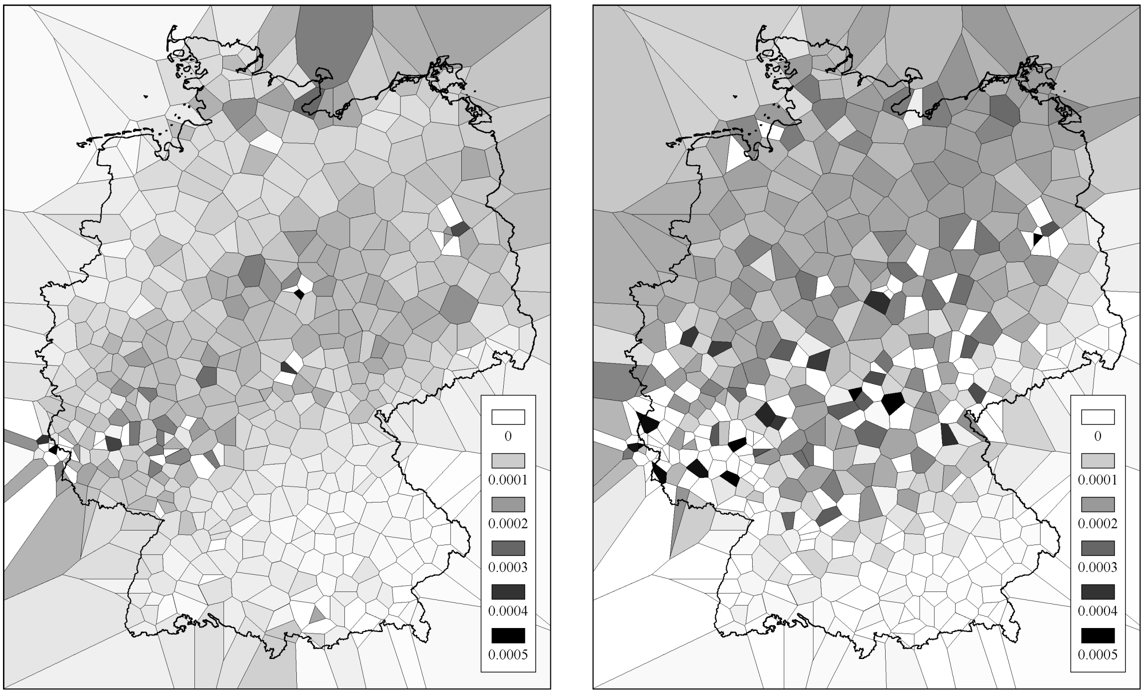
which describes a system of n linear equations with n unknown variables $a_1^{(r')}, \dots, a_n^{(r')}$. There is a unique (exact) solution of (4.6) but in most cases, this solution has one or more negative entries, which contradicts the concept of intensities. Therefore, we suggest to solve (4.6) under the constraint that $a_1^{(r')}, \dots, a_n^{(r')} \geq 0$, which, however, implies that an exact solution does not exist in general. For that reason, we compute $a_1^{(r')}, \dots, a_n^{(r')}$ in a nonnegative least squares sense by

$$\big(a_1^{(r')}, \dots, a_n^{(r')}\big) = \underset{a'_1, \dots, a'_n \geq 0}{\operatorname{argmin}} \left\{ \sum_{j=1}^{n} \left(\log\left(\frac{1}{1-p(s_j)}\right) - \sum_{i=1}^{n} a'_i\, \nu_2\big(b(s_j, r') \cap V(s_i)\big) \right)^2 \right\} \tag{4.7}$$

according to the algorithm given in Lawson and Hanson (1974), Chapter 23. In general, the intensities $a_1^{(r')}, \dots, a_n^{(r')}$ determined according to (4.7) are not an exact solution of (4.5). However, a comparison between the available probabilities $p(s_1), \dots, p(s_n)$ from data and probabilities computed by (4.5) with intensities derived according to (4.7) reveals that the difference is negligible. In particular, no systematic bias is observed. Figure 4.4 illustrates computed intensities for two sample forecast periods with different fixed cell radii. A comparison with Figure 4.2 shows a high accordance of available point probabilities and computed intensities, i.e., higher intensities are obtained in areas with higher probabilities of precipitation and vice versa. The effect of the cell radius r' on computed intensities can be explained as follows. If for a fixed forecast period the radius r' is increased, i.e., precipitation cells cover a wider area, then the mean number of cells in W needs to be decreased accordingly (such that the point probabilities $p(s_1), \dots, p(s_n)$ are still matched), which results in smaller intensities $a_1^{(r')}, \dots, a_n^{(r')}$. This is the reason why computed intensities in central Germany for the forecast period December 9, 2012, 11-12 UTC are only moderately higher compared to the period June 16, 2012, 20-21 UTC although the difference in the corresponding point probabilities is considerably larger. In Sections 4.5.2 and 4.5.3, a method for the computation of the cell radius r is proposed, such that (4.5) can be assumed to hold for all $t \in W$. After r has been determined, the intensities $a_1, \dots, a_n$ are given by $a_i = a_i^{(r)}$ for $i = 1, \dots, n$.

4.5.2. Iterative semivariogram estimation

Intuitively, the radius of precipitation cells should be closely related to the spatial correlation structure of the random field $\{P(t), t \in W\}$ of point probabilities, at least



June 16, 2012, 20-21 UTC, $r' = 30$ pixel December 9, 2012, 11-12 UTC, $r' = 52.5$ pixel

Figure 4.4.: Local intensities for the occurrence of precipitation cells computed for two selected forecast periods with different precipitation cell radii. Each Voronoi cell is colored according to the corresponding local intensity.

for small distances. To quantify the degree of spatial dependence, we suggest to consider the semivariogram $\gamma_P^\star : W \times W \to [0, \infty)$ of $\{P(t), t \in W\}$, see Section 3.2.5. However, a meaningful estimation and analysis of the semivariogram based on realizations of $P(s_1), \ldots, P(s_n)$ is only possible if $\{P(t), t \in W\}$ can be assumed to be second-order motion-invariant. While it is fairly realistic that for $s, t \in W$, the semivariogram $\gamma_P^\star(s, t)$ only depends on the distance $\|s - t\|_2$ when geographical features are ignored, the assumption of $\{P(t), t \in W\}$ having a constant expectation function can generally not be justified. Depending on the spatially varying current weather conditions in the considered forecast period T, the precipitation probabilities show a clear spatial trend, see, e.g., the available data illustrated in Figure 4.2.

A possible approach to address this problem is to decompose the random field $\{P(t), t \in W\}$ into a deterministic expectation function $\{\mu(t), t \in W\}$ (also called trend function) with $\mu(t) = \mathbb{E}\, P(t)$ for $t \in W$ and a random field of residuals $\{\xi(t), t \in W\}$ with $\xi(t) : \Omega \to \mathbb{R}$ and $\mathbb{E}\, \xi(t) = 0$ for $t \in W$, i.e.,

$$P(t) = \mu(t) + \xi(t), \quad t \in W. \tag{4.8}$$

Then, the semivariogram $\gamma_\xi^\star : W \times W \to [0,\infty)$ of the random field $\{\xi(t), t \in W\}$ of residuals is identical to the semivariogram $\gamma_P^\star$ of $\{P(t), t \in W\}$ but now the random field $\{\xi(t), t \in W\}$ can be assumed to be second-order motion-invariant. Therefore, $\{\xi(t), t \in W\}$ has a motion-invariant semivariogram $\gamma_\xi : [0,\infty) \to [0,\infty)$ defined by

$$\gamma_\xi(r) = \gamma_\xi^\star(s,t), \quad s,t \in W \text{ such that } \|s-t\|_2 = r.$$

Furthermore, second-order motion-invariance implies that the variance $\sigma^2 = \operatorname{var} \xi(t)$ does not depend on $t \in W$. However, a reasonable estimation of γ_ξ is a certain problem since only realizations of $P(s_1), \ldots, P(s_n)$ are given, whereas neither $\mu(s_1), \ldots, \mu(s_n)$ nor realizations of $\xi(s_1), \ldots, \xi(s_n)$ are available. Thus, commonly used estimators as defined in Section 3.2.6 are not directly applicable.

As an alternative, we consider an iterative approach proposed in Neuman and Jacobson (1984). Suppose that the expectation function $\{\mu(t), t \in W\}$ can be represented as

$$\mu(t) = \sum_{j=1}^{k} f_j(t)\beta_j, \quad t \in W, \tag{4.9}$$

where $k \in \mathbb{N}$ is an arbitrary integer, $f_1, \ldots, f_k : W \to \mathbb{R}$ is a sequence of base functions, and $\beta_1, \ldots, \beta_k \in \mathbb{R}$ are certain trend coefficients. In our example of application, we put $k = 10$ and $(f_1(t), \ldots, f_{10}(t)) = (1, t_{(1)}, t_{(2)}, t_{(1)}^2, t_{(1)}t_{(2)}, t_{(2)}^2, t_{(1)}^3, t_{(1)}^2 t_{(2)}, t_{(1)} t_{(2)}^2, t_{(2)}^3)$ are monomials ranging up to 3rd order for any location $t = (t_{(1)}, t_{(2)}) \in W$. We introduce the following simplifying notation. Let $\boldsymbol{P} = (P(s_1), \ldots, P(s_n))^\top$ be the random vector of point probabilities at $s_1, \ldots, s_n$, $\boldsymbol{X} = (x_{ij}) \in \mathbb{R}^{n \times k}$ with $x_{ij} = f_j(s_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$ a deterministic design matrix, $\boldsymbol{\xi} = (\xi(s_1), \ldots, \xi(s_n))^\top$ the random vector of residuals at $s_1, \ldots, s_n$, $\Sigma \in \mathbb{R}^{n \times n}$ the covariance matrix of $\boldsymbol{\xi}$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ a trend vector. Combining (4.8) and (4.9) with the notation introduced above yields that

$$\boldsymbol{P} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi}, \tag{4.10}$$

which is a linear regression model with correlated residuals. Since the covariance matrix Σ of $\boldsymbol{\xi}$ is unknown, GLS cannot be used for the estimation of the trend vector $\boldsymbol{\beta}$. Instead, OLS could be applied, but this would result in strongly biased estimators for $\boldsymbol{\beta}$ and Σ. The approach presented in the following provides (iterative) estimators of $\boldsymbol{\beta}$ and Σ with drastically reduced biases, although this bias cannot be removed completely. This is due to the fact that the estimation of the covariance matrix Σ of $\boldsymbol{\xi}$ based on empirical residuals obtained from the GLS estimator is always biased in linear regression with correlated residuals, even if Σ is known. In Cressie (1993), Chapter 3.4.3 and Beckers and Bogaert (1998), some simple examples are given to illustrate this bias problem.

Based on the assumptions made above, estimators of the trend vector $\boldsymbol{\beta}$, the covariance matrix Σ, and the (motion-invariant) semivariogram γ_ξ of $\{\xi(t), t \in W\}$ can be computed according to an iterative algorithm proposed in Neuman and Jacobson (1984). To allow for a quick and stable convergence of this algorithm, it is further assumed that γ_ξ is an exponential semivariogram with parameter vector $\theta = (c_0, c, a_e)$ where $c_0 \geq 0$ and $c, a_e > 0$, see Example 3.2.2. To emphasize this, we write γ_ξ^θ instead of γ_ξ in the following.

1. Compute an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ using OLS according to

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{b} \in \mathbb{R}^k}{\operatorname{argmin}} \left\{ (\boldsymbol{P} - \boldsymbol{X}\boldsymbol{b})^\top (\boldsymbol{P} - \boldsymbol{X}\boldsymbol{b}) \right\} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}.$$

2. Use the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ obtained in the previous step to determine a vector $\hat{\boldsymbol{\xi}}$ of empirical residuals according to

$$\hat{\boldsymbol{\xi}} = \boldsymbol{P} - \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

   Based on $\hat{\boldsymbol{\xi}}$, compute a method of moment estimator of the semivariogram γ_ξ^θ according to (3.17) and determine an estimator $\hat{\theta}$ of the parameter vector θ using the iterative algorithm presented in Section 3.2.7. Furthermore, compute a method of moment estimator $\hat{\sigma}^2$ of the variance σ^2 based on $\hat{\boldsymbol{\xi}}$. Finally, determine a plug-in estimator $\hat{\Sigma} = (\hat{\sigma}_{ij}^2)_{i,j=1,\ldots,n}$ of the covariance matrix Σ of $\boldsymbol{\xi}$ using the unique relationship between the semivariogramm and the covariance function of a random field, i.e., by

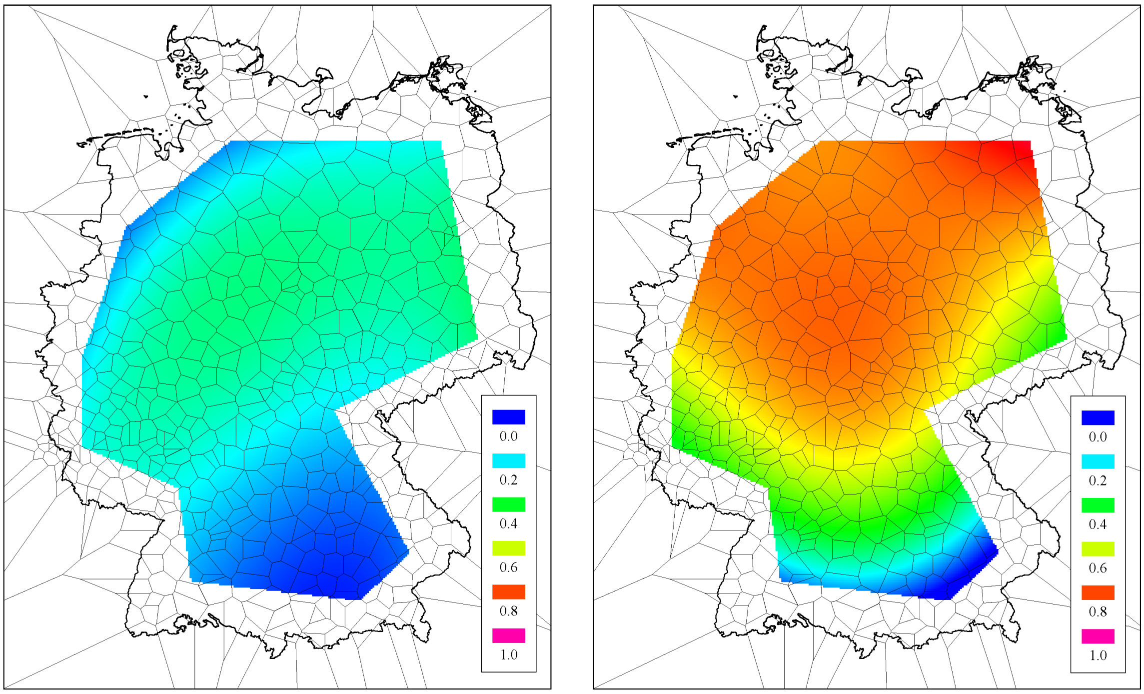
$$\hat{\sigma}_{ij}^2 = \hat{\sigma}^2 - \gamma_\xi^{\hat{\theta}}(\|s_i - s_j\|_2), \quad i, j = 1, \ldots, n. \tag{4.11}$$

3. Recompute the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ using GLS with (estimated) covariance matrix $\hat{\Sigma}$, i.e., by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{b} \in \mathbb{R}^k}{\operatorname{argmin}} \left\{ (\boldsymbol{P} - \boldsymbol{X}\boldsymbol{b})^\top \hat{\Sigma}^{-1} (\boldsymbol{P} - \boldsymbol{X}\boldsymbol{b}) \right\} = (\boldsymbol{X}^\top \hat{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \hat{\Sigma}^{-1} \boldsymbol{P}.$$

4. Repeat steps 2 and 3 until $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\xi}}$, and $\hat{\theta}$ converge to stable values. Once the algorithm is terminated, $\hat{\gamma}$ denotes the estimator of γ_ξ^θ that is obtained according to the exponential semivariogram model with estimated parameter vector $\hat{\theta}$.

Fitting a semivariogram model in step 2 is necessary to allow for a stable and quick convergence of the algorithm (which can not be guaranteed if a method of moment estimator of the semivariogram is used in (4.11)). Comparisons of different semivariogram models for a sequence of sample forecast periods have shown that an exponential semivariogram seems to provide the best overall fit. Furthermore, note that the estimator



June 16, 2012, 20-21 UTC December 9, 2012, 11-12 UTC

Figure 4.5.: Estimated trend functions for random point probabilities of precipitation in W_{in} for two selected forecast periods.

$\hat{\Sigma}$ specified in (4.11) does not correspond to a method of moment estimator (even if the fitting of the exponential semivariogram model is skipped), see Remark 3.2.12. Figure 4.5 illustrates estimated trend functions of random point probabilities that are computed according to (4.9) for two sample forecast periods. In order to avoid boundary effects, we only consider values of the trend functions in a compact subset W_{in} of W, which contains all locations with a certain distance to the boundaries of Germany. A comparison to the available data depicted in Figure 4.2 shows a good accordance. Examples of estimated semivariograms are illustrated in Section 4.5.3.

4.5.3. Radius of precipitation cells

To conclude the computation of model characteristics, we describe how the radius R of precipitation cells can be determined based on estimated semivariograms. Again, we consider a fixed realization e of the random error E and the corresponding realizations $p(s_1) = \mathbb{E}\,(P(s_1)\,|\,E = e), \ldots, p(s_n) = \mathbb{E}\,(P(s_n)\,|\,E = e)$, $a_1 = \mathbb{E}\,(A_1\,|\,E = e), \ldots, a_n = \mathbb{E}\,(A_n\,|\,E = e)$, and $r = \mathbb{E}\,(R\,|\,E = e)$ of the random point probabilities

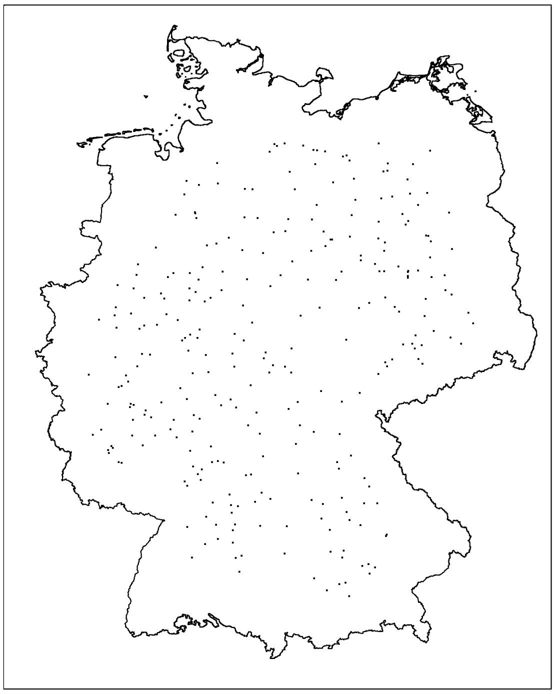


Figure 4.6.: Locations of estimation points used for the computation of point probabilities in the context of determining the radius of precipitation cells.

$P(s_1), \ldots, P(s_n)$, the random local intensities $A_1, \ldots, A_n$ for the occurrence of precipitation cells, and the random cell radius R. We aim to configure model characteristics in such a way that computed point probabilities from the model and those from the data have a similar dependency structure. For that purpose, we first compute an estimator $\hat{\gamma}$ for the semivariogram of the random field of residuals that corresponds to $\{P(t), t \in W\}$ based on $p(s_1), \ldots, p(s_n)$ using the iterative algorithm introduced in Section 4.5.2. For comparison, similar semivariograms are estimated based on point probabilities that are derived using various model configurations.

We consider a sequence $t_1, \ldots, t_l$ of $l \in \mathbb{N}$ locations in W (denoted as estimation points in the following) that are comparable to (but different from) $s_1, \ldots, s_n$. In our example of application, estimation points are chosen as a realization of a stationary Poisson point process inside the window W_{in} (as point probabilities can only be determined reliably for locations in W_{in}), which is generated according to Algorithm 3.3.2. The intensity of the process is chosen such that the mean number of estimation points in W_{in} is equal to $\#\{i \in \{1, \ldots, n\} : s_i \in W_{in}\}$. The set of estimation points used to obtain the results that are illustrated in the following is depicted in Figure 4.6. For any possible radius $r' > 0$, we introduce the deterministic field $\{p^{(r')}(t), t \in W\}$ of point precipitation probabilities that are computed as coverage probabilities of the germ-grain model M of precipitation cells with radius $r = r'$ and local intensities

$a_i = a_i^{(r')}$ for $i = 1, \ldots, n$ by

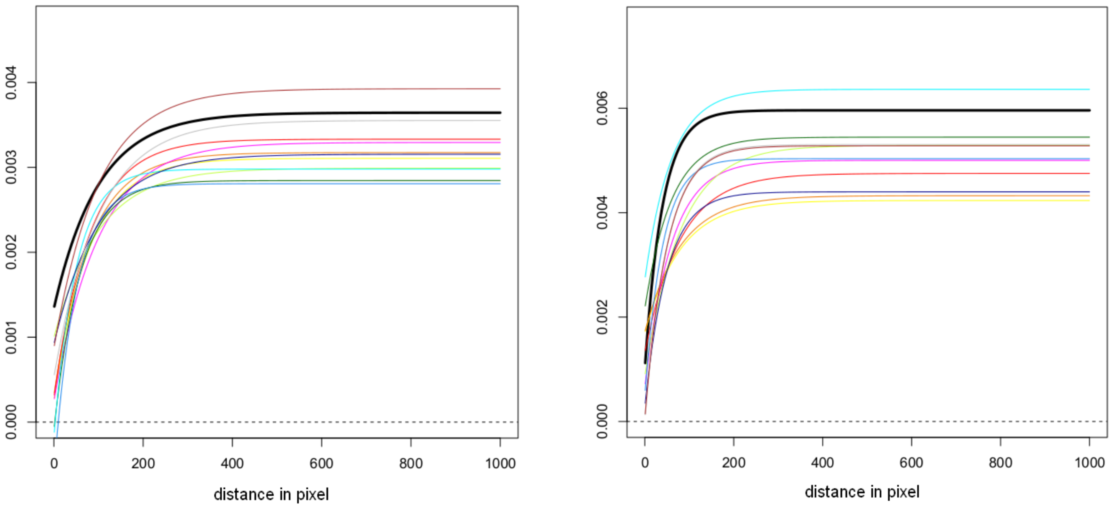
$$p^{(r')}(t) = 1 - \exp\left\{ -\sum_{i=1}^{n} a_i^{(r')}\, \nu_2\big(b(t, r') \cap V(s_i)\big) \right\}, \quad t \in W,$$

where $a_1^{(r')}, \ldots, a_n^{(r')}$ are determined according to (4.7). We suppose that $\{p^{(r')}(t), t \in W\}$ is a realization of some random field $\{P^{(r')}(t), t \in W\}$ such that the assumptions made in Section 4.5.2 (in particular (4.8) and (4.9)) remain valid if $\{P(t), t \in W\}$ is replaced by $\{P^{(r')}(t), t \in W\}$. According to this, we consider the motion-invariant semivariogram $\gamma_\xi^{(r')} : [0, \infty) \to [0, \infty)$ of the (second-order motion-invariant) random field of residuals $\{\xi^{(r')}(t), t \in W\}$ that corresponds to $\{P^{(r')}(t), t \in W\}$ and determine an estimator $\hat{\gamma}^{(r')}$ of $\gamma_\xi^{(r')}$ based on $p^{(r')}(t_1), \ldots, p^{(r')}(t_l)$ using the iterative algorithm presented in Section 4.5.2. Finally, we choose the radius r of precipitation cells in such a way that the squared $\mathcal{L}_2$-distance between $\hat{\gamma}$ and $\hat{\gamma}^{(r)}$ is minimal among all estimated semivariograms $\{\hat{\gamma}^{(r')}, r' > 0\}$, i.e.,

$$r = \underset{r'>0}{\operatorname{argmin}} \left\{ \int_{c_1}^{c_2} \left(\hat{\gamma}(h) - \hat{\gamma}^{(r')}(h) \right)^2 \mathrm{d}h \right\} \tag{4.12}$$

with $c_1, c_2 > 0$ being some suitable integration limits. In practice, a finite set $\{r'_1, \ldots, r'_k\}$ of possible radii is considered in (4.12), which we suggest to be chosen according to available computing power. We furthermore recommend to provide a lower limit for the cell radius r that depends on the sizes of the Voronoi cells $V(s_1), \ldots, V(s_n)$. The smaller the radius r is, the more Voronoi cells $V(s_i)$ satisfy that $\nu_2\big(b(t, r) \cap V(s_i)\big) = 0$ for a fixed $t \in W$, making the field of obtained point probabilities resemble a piecewise constant function if r is too small. In contrast, larger values of r lead to smoother fields of point probabilities.

Finally, we illustrate the results which are obtained in our example of application. In Figure 4.7, the estimated semivariogram $\hat{\gamma}$ (black) is compared to the estimated residual semivariograms $\hat{\gamma}^{(r')}$ that correspond to a sequence of 15 possible cell radii (in colors) for two selected sample forecast periods. To account for the observation that typically precipitation patterns have smaller scales in summer than in winter periods, we consider the possible radii $20, 22.5, 25, \ldots, 55$ (in pixel) for all forecast periods in summer 2012 and the radii $25, 27.5, 30, \ldots, 60$ for winter 2012. The radius r obtained by the proposed method is $r = 30$ for forecast period June 16, 2012, 20-21 UTC and $r = 52.5$ for forecast period December 9, 2012, 11-12 UTC. It is clearly visible in Figure 4.7 that the radius of modeled precipitation cells indeed has an impact on the dependency structure of the residuals and that the semivariogram estimated from the data goes well with those estimated for different model configurations. For the majority of forecast periods we obtain similar results.



June 16, 2012, 20-21 UTC, possible radii: $20, 22.5, 25, \ldots, 55$ pixel.

December 9, 2012, 11-12 UTC, possible radii: $25, 27.5, 30, \ldots, 60$ pixel.

Figure 4.7.: Comparison of semivariograms estimated based on available data (black) and based on the introduced model of precipitation cells for different cell radii (in colors) using the iterative algorithm presented in Section 4.5.2.

4.6. Model-based computation and estimation of area probabilities

After suitable statistical methods have been developed for the computation of the model characteristics $A_1, \ldots, A_n$ and R based on available point probabilities $P(s_1), \ldots, P(s_n)$, the germ-grain model M of precipitation cells introduced in (4.3) can now be used for the computation of area probabilities. Similar to the considerations made when modeling point probabilities in Section 4.3, we suppose that there is precipitation somewhere inside a Borel set $B \in \mathcal{B}(W)$ if B intersects at least one grain of the random union set M (i.e., one precipitation cell). Accordingly, the following representation formula for the random area probability $\Pi(B) : \Omega \to [0, 1]$ of B can be derived using the distributional properties of the Poisson process stated in (3.23):

$$\begin{aligned}\Pi(B) &= \mathbb{P}(B \cap M \neq \emptyset \,|\, E) \\ &= 1 - \mathbb{P}\big(\#\{i : X_i \in B \oplus b(o, R)\} = 0 \,|\, E\big) \\ &= 1 - \exp\left\{-\int_{B \oplus b(o,R)} \sum_{i=1}^{n} A_i \,\mathbb{1}_{V(s_i)}(z)\,\mathrm{d}z\right\}\end{aligned}$$

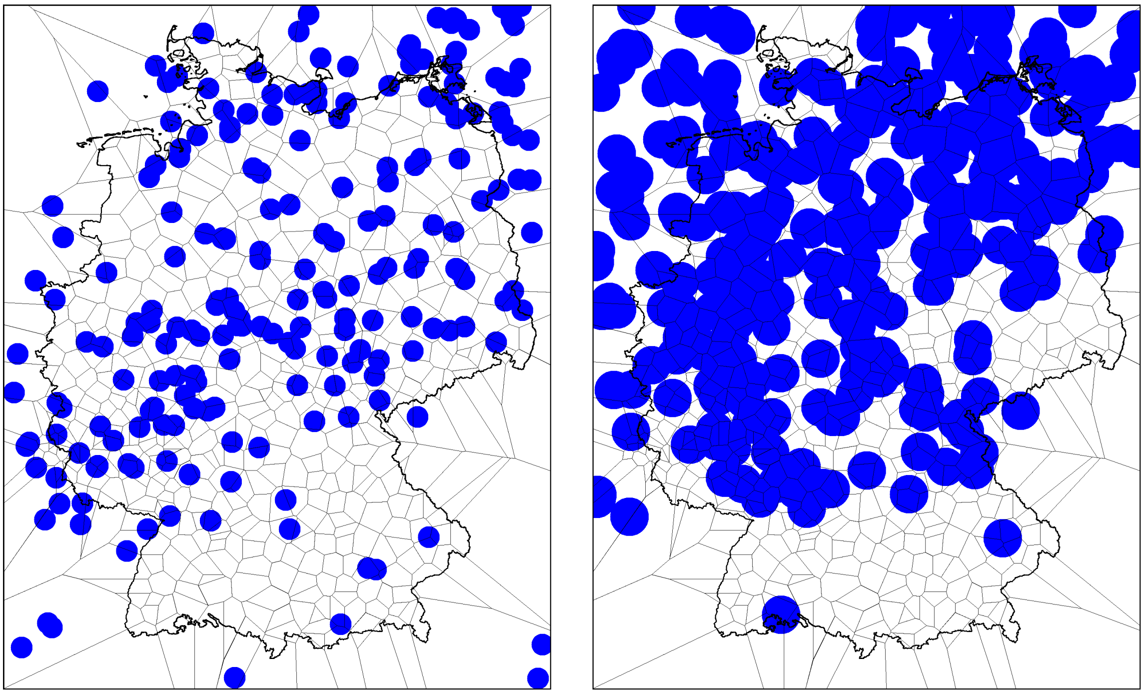
$$= 1 - \exp\left\{ - \sum_{i=1}^{n} A_i \int_W \mathbb{1}_{(B \oplus b(o,R)) \cap V(s_i)}(z)\, \mathrm{d}z \right\}$$
$$= 1 - \exp\left\{ - \sum_{i=1}^{n} A_i\, \nu_2\big((B \oplus b(o,R)) \cap V(s_i)\big) \right\}. \tag{4.13}$$

This representation indeed is a generalization of the formula for point probabilities derived in Section 4.4 since for $B = \{t\}$ with $t \in W$, (4.13) is identical to (4.4). When using the proposed model in applications (with a fixed realization e of E), the model characteristics $a_1, \ldots, a_n$ and r are first determined based on the sequence $p(s_1), \ldots, p(s_n)$ of point probabilities according to (4.7) with $r' = r$ and (4.12). Then, the (deterministic) area probability $\pi(B) = \mathbb{E}\,(\Pi(B) \mid E = e)$ of each Borel set $B \in \mathcal{B}(W)$ can be computed based on $a_1, \ldots, a_n$ and r using the representation formula in (4.13). Alternatively, $\pi(B)$ can also be estimated using repeated simulation of the germ-grain model M. Recall that M is a Boolean model conditioned on $\{E = e\}$ that can be generated efficiently using Algorithm 3.3.5, which suggests to use the following estimator of $\pi(B)$. For an arbitrary integer $j \in \mathbb{N}$ let $M_1, \ldots, M_j$ be a sequence of (conditionally) independent and identically distributed Boolean models with $M_i \overset{\mathrm{d}}{=} M$ for $i = 1, \ldots, j$ (conditioned on $\{E = e\}$). Then, a Monte Carlo estimator $\hat{\pi}(B)$ of $\pi(B)$ is given by

$$\hat{\pi}(B) = \frac{1}{j}\, \#\{i \in \{1, \ldots, j\} : B \cap M_i \neq \emptyset\}. \tag{4.14}$$

It depends on the application whether area probabilities should be computed directly according to the model or estimated based on simulations. In general, using (4.13) has the disadvantage that the intersection areas of the dilated set $B \oplus b(o, r)$ and the Voronoi cells $V(s_1), \ldots, V(s_n)$ need to be determined numerically, which is both computationally expensive and imprecise. This suggests that using the Monte Carlo approach might be favorable in many situations. On the other hand, if the model is applied in operational weather forecasting on a daily base (or even more frequently) for the computation of area probabilities of a fixed Borel set $B \in \mathcal{B}(W)$, it might be more efficient to determine the intersection areas $\nu_2((B \oplus b(o, r)) \cap V(s_i))$ for $i = 1, \ldots, n$ once for all possible cell radii, which then allows to quickly compute area probabilities using the direct formula for different forecast periods.

To conclude this section, we present some results that were obtained in our example of application. Figure 4.8 shows typical realizations of the germ-grain model M for two sample forecast periods. We observe that although precipitation cells look different from precipitation patterns in radar data, see Figure 4.3, they mainly occur in the same regions. Furthermore, a close correspondence to the underlying point probabilities and the computed intensities for the occurrence of precipitation cells is found, compare to Figures 4.2 and 4.4. Finally, examples of point and area probabilities are illustrated that were estimated based on 2,000 realizations of M. To avoid boundary effects, only



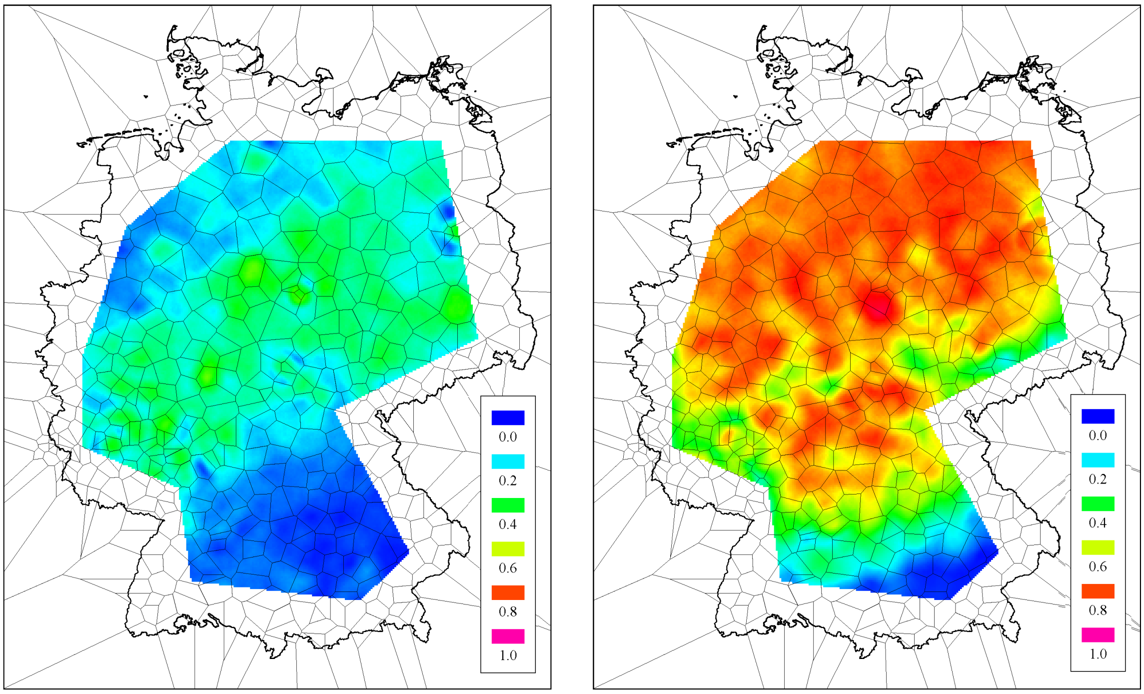
June 16, 2012, 20-21 UTC, $r = 30$ pixel December 9, 2012, 11-12 UTC, $r = 52.5$ pixel

Figure 4.8.: Realizations of the germ-grain model M of precipitation cells with characteristics computed from available point probabilities for two selected forecast periods.

probabilities for locations in W_{in} and areas intersecting W_{in} are considered. Figure 4.9 shows estimated point probabilities for two sample forecast periods, which correspond well to the point probabilities from the data, compare to Figure 4.2. In particular, smooth transitions of point probabilities at the boundaries of the Voronoi cells are provided. In Figure 4.10, area probabilities for all Voronoi cells $V(s_1), \ldots, V(s_{503})$ that intersect W_{in} are depicted, where each Voronoi cell is colored according to the corresponding area probability. We see that obtained area probabilities are clearly higher than the corresponding point probabilities but are generally consistent providing high values in regions with higher point probabilities and vice versa.

4.7. Forecast verification

In order to assess whether the proposed model of precipitation cells is able to provide precise and reliable area probabilities, we perform a verification of forecasts using precipitation analyses derived from rain gauge adjusted radar data (RADOLAN), see

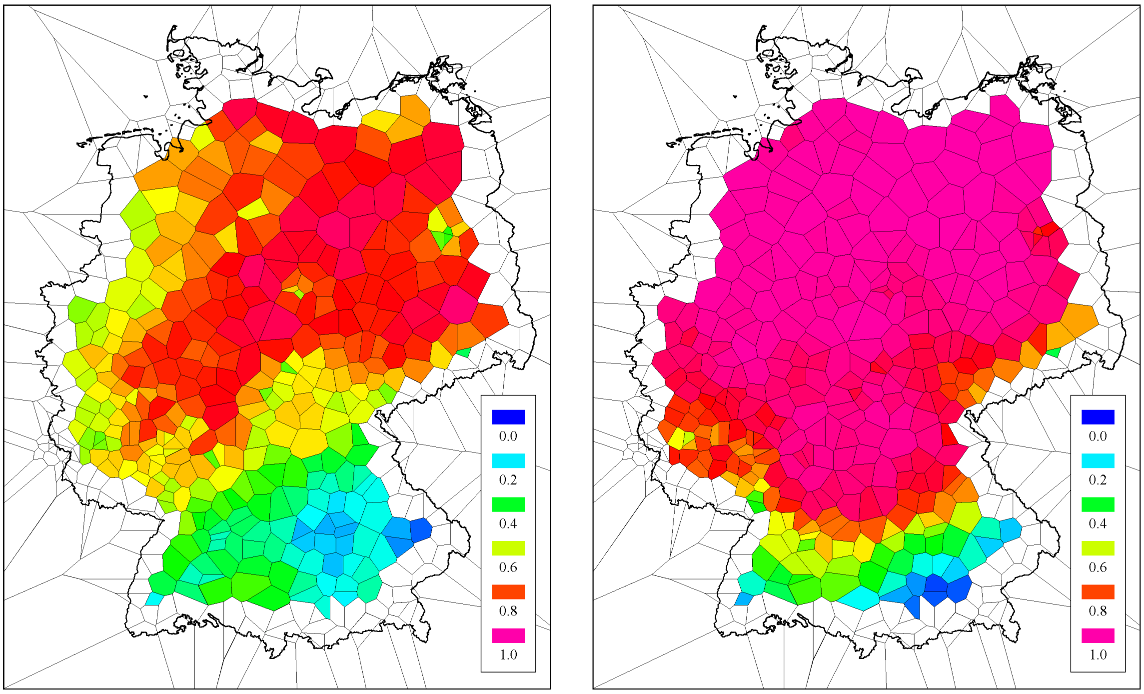


June 16, 2012, 20-21 UTC December 9, 2012, 11-12 UTC

Figure 4.9.: Point probabilities estimated for locations in W_{in} based on 2,000 realizations of the germ-grain model M of precipitation cells for two selected forecast periods.

Section 4.1. A comparison of Figures 4.3 and 4.10 already suggests a high correspondence of provided area probabilities and radar data (for two considered forecast periods) as in the vast majority of Voronoi cells with high area probabilities precipitation indeed occurred and in most regions of low probabilities no precipitation was observed. Now, the goal of forecast verification is to quantify the relationship between area probabilities and radar data numerically over the entire sample period. For that purpose, we use score functions and diagrams that are commonly used by meteorologists to assess several aspects of forecast performance. However, note that a variety of other verification tools exist that are tailored for different weather events and forecast types. A more detailed overview of this topic can be found, e.g., in Wilks (2011), Section 8.4.

For any test area $B \in \mathcal{B}(W)$ that intersects W_{in}, the derived direct representation formula (4.13) or the Monte Carlo estimator (4.14) can be used to determine area probabilities $\pi_1(B), \ldots, \pi_m(B)$ that correspond to the $m = 854$ available forecast periods described in Section 4.1. Furthermore, a sequence of precipitation indicators $I_1(B), \ldots, I_m(B)$ is considered, where $I_j(B)$ is equal to 1 if there is precipitation



June 16, 2012, 20-21 UTC December 9, 2012, 11-12 UTC

Figure 4.10.: Area probabilities estimated for Voronoi cells that intersect W_{in} based on 2,000 realizations of the germ-grain model M of precipitation cells for two selected forecast periods.

somewhere within B in forecast period $j \in \{1, \ldots, m\}$ with respect to RADOLAN data and 0 otherwise.

4.7.1. Bias, Brier skill score, and empirical correlation coefficient

We consider the following three score functions that provide a systematic comparison of area probabilities and precipitation indicators. One of the most commonly used scores is the bias b_π, which is defined as the difference between the mean precipitation probability of B and the relative frequency of precipitation occurring in B over the sample period, i.e.,

$$b_\pi = \frac{1}{m}\sum_{j=1}^{m} \pi_j(B) - \frac{1}{m}\sum_{j=1}^{m} I_j(B). \tag{4.15}$$

For a good weather prediction, the bias b_π given in (4.15) should be as close to zero as possible. A clearly negative or positive bias would indicate that computed probabilities are systematically too low or too high, respectively. However, neither does an unbiased forecast necessarily have a high quality nor does a (small) bias always imply bad forecast performance. For example, the climate mean, which provides area probabilities $\pi'_1(B), \dots, \pi'_m(B)$ according to

$$\pi'_i(B) = \frac{1}{m} \sum_{j=1}^{m} I_j(B), \quad i = 1, \dots, m, \tag{4.16}$$

has a perfect bias of zero but is clearly not a good forecast as for each forecast period the same area probability is obtained. Furthermore, note that the climate mean as defined in (4.16) is only based on past forecast periods and is thus a poor choice in operational weather prediction.

It is advised to rely on more than one score function when verifying weather forecasts that address different aspects of forecast quality, see Wilks (2011), Section 8.1. A popular measure of accuracy for probabilistic forecasts is the Brier score (BS) bs_π (Brier, 1950), which is defined as the mean squared difference of predicted probabilities and observed precipitation indicators, i.e.,

$$bs_\pi = \frac{1}{m} \sum_{j=1}^{m} \big(\pi_j(B) - I_j(B)\big)^2. \tag{4.17}$$

Obviously, the BS should be as small as possible. A score of $bs_\pi = 0$ is achieved if $\pi_j(B) = I_j(B)$ for all $j = 1, \dots, m$, i.e., if the occurrence or non-occurrence of precipitation is always forecasted correctly with probability 1. Conversely, the worst forecast quality is obtained when $bs_\pi = 1$, i.e., if precipitation is always forecasted incorrectly with probability 1 (in fact, $bs_\pi \in \{0, 1\}$ rarely occurs in applications since probabilities of precipitation typically have values in $(0, 1)$). However, a more intuitive interpretation of this score is difficult since the BS also depends on a term denoted as uncertainty, which is expressed only by the variability of $I_1(B), \dots, I_m(B)$, see, e.g., Wilks (2011), Section 8.4.3. Therefore, it is common to consider a score function that relates the BS of the area probabilities $\pi_1(B), \dots, \pi_m(B)$ to the BS of a reference prediction. Assume that another sequence $\tilde{\pi}_1(B), \dots \tilde{\pi}_m(B)$ of area precipitation probabilities computed from a different method is available and consider the BS $bs_{\tilde{\pi}}$, which is determined according to (4.17) using $\tilde{\pi}_1(B), \dots \tilde{\pi}_m(B)$ instead of $\pi_1(B), \dots, \pi_m(B)$. If no reference method for the derivation of area probabilities is available, which is the case in this thesis, then the climate mean defined in (4.16) is used, i.e., we set $\tilde{\pi}_i(B) = \pi'_i(B)$ for $i = 1, \dots, m$. Finally, the so-called Brier skill score (BSS) bss_π defined as

$$bss_\pi = 1 - \frac{bs_\pi}{bs_{\tilde{\pi}}} \tag{4.18}$$

is computed in order to investigate whether the considered forecasts have a higher quality (in terms of BSs) than the reference method. If this is the case, the BSS should be clearly positive. In contrast, negative values of the BSS indicate that the verified forecasts perform worse than the reference method (which is particularly bad if the reference method is, as in our case, given by the climate mean).

As a third score function we suggest to consider the empirical correlation coefficient $\hat{\rho}$ of the computed probabilities $\pi_1(B), \ldots, \pi_m(B)$ and the observed precipitation indicators $I_1(B), \ldots, I_m(B)$ given as

$$\hat{\rho} = \frac{\sum_{j=1}^{m} (\pi_j(B) - \hat{\mu}_\pi)(I_j(B) - \hat{\mu}_I)}{\sqrt{\sum_{j=1}^{m} (\pi_j(B) - \hat{\mu}_\pi)^2 \sum_{j=1}^{m} (I_j(B) - \hat{\mu}_I)^2}}, \tag{4.19}$$

with

$$\hat{\mu}_\pi = \frac{1}{m} \sum_{j=1}^{m} \pi_j(B) \quad \text{and} \quad \hat{\mu}_I = \frac{1}{m} \sum_{j=1}^{m} I_j(B).$$

Clearly, $\hat{\rho}$ should be as close to one as possible and $\hat{\rho} = 0$ indicates that there is no relationship between computed precipitation probabilities and observed precipitation events. Negative correlation coefficients suggest that the provided forecasts should clearly not be considered in applications, as this indicates that on average lower area probabilities are forecasted for B in periods where precipitation occurs than in periods where it does not.

In order to perform a systematic analysis of the three proposed score functions, an adequate set of test areas needs to be chosen. For that purpose, we suggest to consider the Voronoi cells $V(s_1), \ldots, V(s_{503})$ around the locations $s_1, \ldots, s_{503}$ of the weather stations illustrated in Figure 4.1. The Voronoi cells $V(s_1), \ldots, V(s_{503})$ appear to be particularly suitable for forecast verification since they cover a large variety of areas with different sizes, shapes, and orientations. Area probabilities computed according to the germ-grain model of precipitation cells are a function of model characteristics, which in turn are computed from the available point probabilities. Therefore, verification is not only performed for area probabilities of the 503 Voronoi cells $V(s_1), \ldots, V(s_{503})$ but also for point precipitation probabilities from the available data that correspond to the locations $s_1, \ldots, s_{503}$ of weather stations. For that purpose, the bias, BSS, and empirical correlation coefficient are defined analogously to (4.15)-(4.19) by replacing the area probabilities and precipitation indicators by their respective point counterparts.

In Figure 4.11, the biases of given point probabilities from data and computed area probabilities are illustrated. On the left-hand side, each Voronoi cell is colored according to the bias of the point probabilities at the location of the corresponding weather station, whereas on the right-hand side, the bias of the area probabilities for each

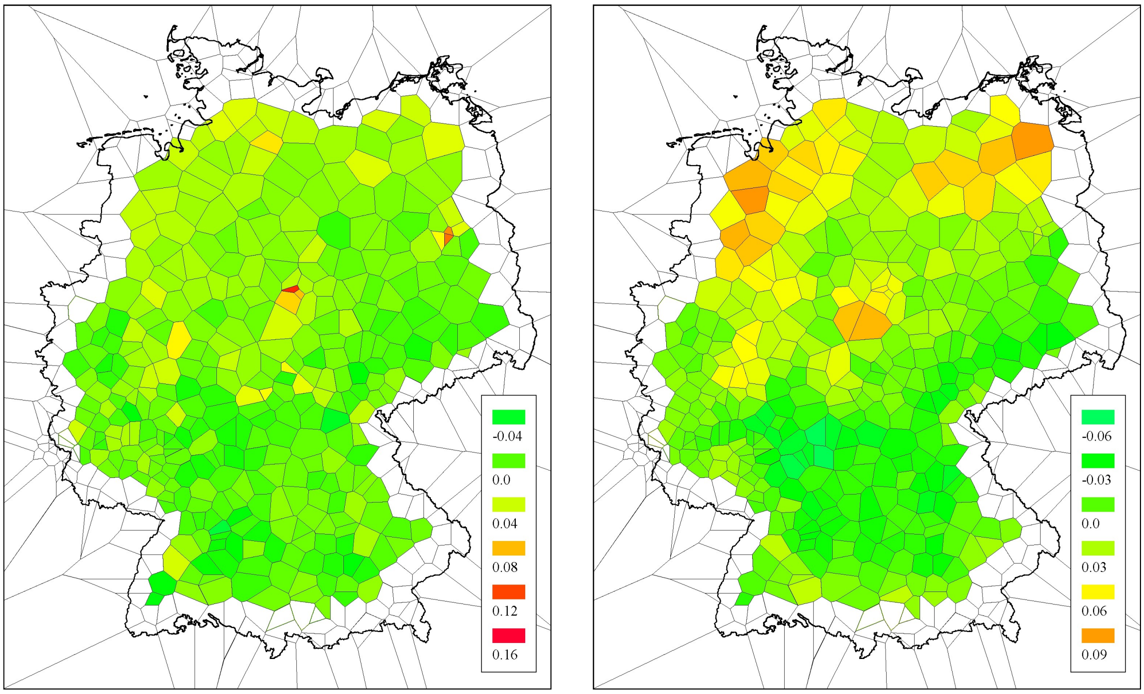


Figure 4.11.: Biases of point probabilities from available data for weather stations $s_1, \ldots, s_{503}$ (left) and of area probabilities computed according to the proposed model of precipitation cells for Voronoi cells $V(s_1), \ldots, V(s_{503})$ (right). Biases are only shown for Voronoi cells intersecting W_{in}.

Voronoi cell is depicted. Again, only cells intersecting W_{in} are considered to avoid boundary effects. The biases for point probabilities range from $-4\,\%$ to $6\,\%$ for most weather stations, which is acceptably small showing a good consistency of the point forecast data with the radar observations. There are a few outliers with biases up to $14\,\%$, which are most likely caused by problems with the radar measurements at these locations. In general, however, no systematic deviation of point probabilities and radar data can be found. The biases for area probabilities range from $-6\,\%$ to $9\,\%$, which is only slightly different from point probabilities. For most test areas no systematic bias is found and the mean bias (over all areas) is at $1\,\%$, which is comparable to the mean bias of point probabilities (over all weather stations). In the north-east and the north-west of Germany, some larger biases are observed indicating that area probabilities are slightly too high, whereas in some regions in southern Germany area probabilities appear to be marginally too low. A similar trend (although to a much smaller extent) is also observed for the point probabilities, which suggests that biases of point probabilities are amplified when deriving area forecasts. Thus, unbiasedness of the underlying data is found to be crucial to obtain reliable area probabilities.

Figure 4.12 provides histograms of BSSs computed for point and area probabilities

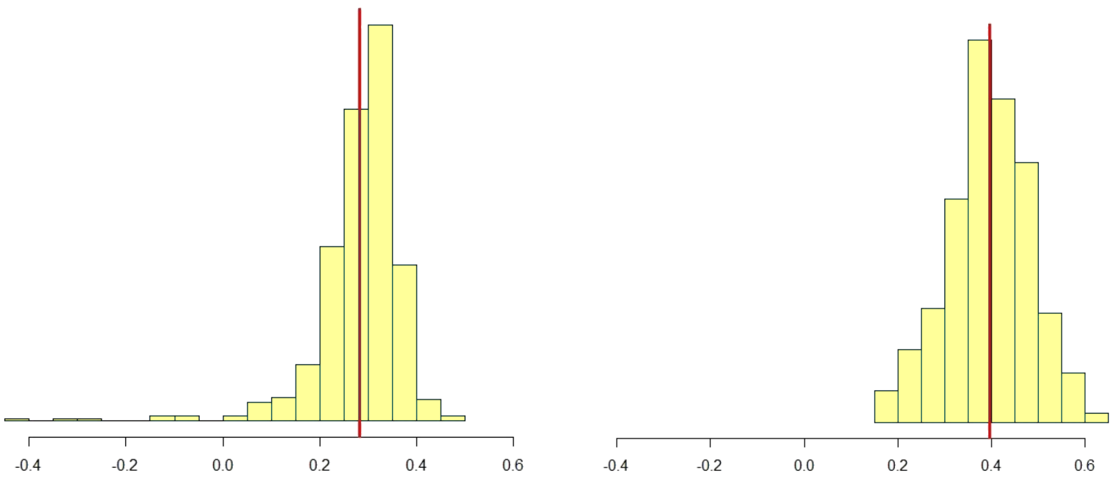


Figure 4.12.: Histograms showing BSSs of point probabilities from available data for weather stations $s_1, \ldots, s_{503}$ (left) and of area probabilities computed according to the proposed model of precipitation cells for Voronoi cells $V(s_1), \ldots, V(s_{503})$ (right). Red lines indicate mean BSSs.

using the climate mean as reference prediction. For the point probabilities of most weather stations positive BSSs are found ranging up to a value of 0.5. There are again outliers for some stations, the mean value of 0.28 shows a positive signal though. The BSS is clearly positive for all considered Voronoi cells, ranging from 0.18 up to more than 0.6. The mean BSS has a value of 0.4 and is thus clearly higher than for point probabilities. This demonstrates that the presented method indeed provides a significant improvement in computing area precipitation probabilities over using the climate mean. Furthermore, it is a pleasant result that the accuracy of area probabilities is not affected by occasional outliers occurring in point forecast data. Similar results are obtained when considering histograms of the empirical correlation coefficients of precipitation probabilities and precipitation indicators from radar data, see Figure 4.13. For all considered weather stations a positive correlation coefficient of the corresponding point probabilities with precipitation indicators is found, although about 20 % of values are smaller than 0.5. The mean correlation coefficient is at 0.54. The histogram on the right shows clearly positive correlations between computed area probabilities and observed precipitation events, too. All values are greater than 0.5 ranging up to more than 0.8 and the mean correlation of 0.65 is significantly higher than that observed for point probabilities. Considering that the correlation coefficient compares continuous probabilities with binary precipitation indicators, this is an excellent result. Furthermore, we point out that our area probabilities perform even better (in terms of BSS and empirical correlation coefficient) than the underlying point data.

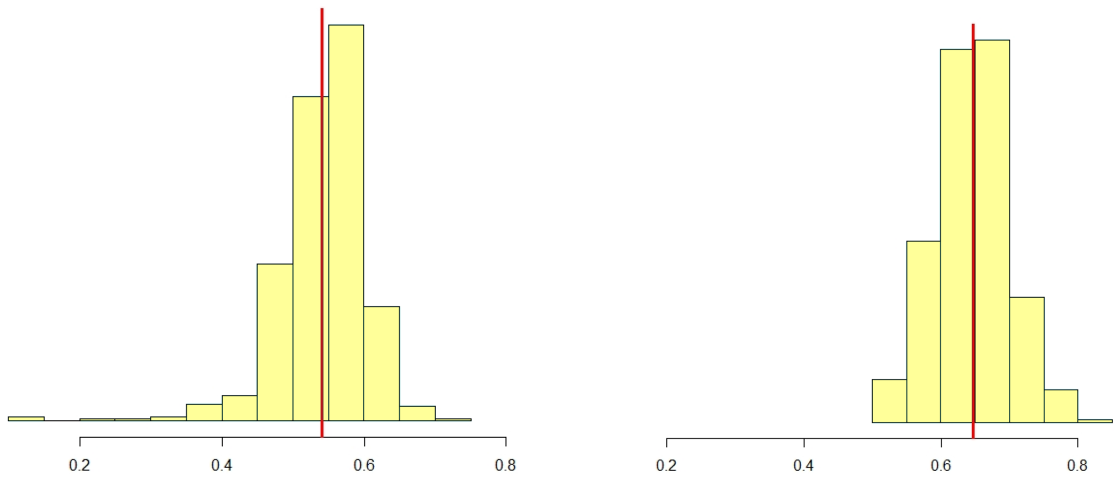


Figure 4.13.: Histograms showing empirical correlation coefficients of point probabilities from available data for weather stations $s_1, \ldots, s_{503}$ (left) and of area probabilities computed according to the proposed model of precipitation cells for Voronoi cells $V(s_1), \ldots, V(s_{503})$ (right) with corresponding precipitation indicators. Red lines indicate mean correlation coefficients.

4.7.2. Reliability diagram

To conclude our validation, reliability diagrams for area precipitation probabilities are considered. Reliability diagrams give a much more detailed view on forecast performance as they aim to describe the joint distribution of forecasts and observations rather than providing single-number summaries as the score functions considered in Section 4.7.1, see Wilks (2011), Section 8.4.4. Basically, a reliability diagram illustrates the empirical conditional distribution of the precipitation indicators given the forecasted area probabilities, i.e., it shows whether precipitation indeed occurs rarely or often in forecast periods with low or high precipitation probabilities, respectively. A mathematical description of the reliability diagram for test area B is given as follows. At first, the unit interval $[0, 1]$ is decomposed into a sequence $U_1, \ldots, U_{20}$ of 20 equal subintervals each having a length of 0.05. Then, for each $k \in \{1, \ldots, 20\}$, the reliability $\hat{\rho}(U_k)$ of the subinterval U_k is defined as the relative frequency of precipitation among all those forecast periods for that the computed area precipitation probability of B takes a value in U_k, i.e.,

$$\hat{\rho}(U_k) = \frac{\#\{j \in \{1, \ldots, m\} : \pi_j(B) \in U_k, I_j(B) = 1\}}{\#\{j \in \{1, \ldots, m\} : \pi_j(B) \in U_k\}}, \quad k = 1, \ldots, 20. \tag{4.20}$$

Additionally, the midpoints $m_1, \ldots, m_{20}$ of the intervals $U_1, \ldots, U_{20}$ are computed and the sequence of points $(m_1, \hat{\rho}(U_1)), \ldots, (m_{20}, \hat{\rho}(U_{20}))$ is called a reliability diagram.

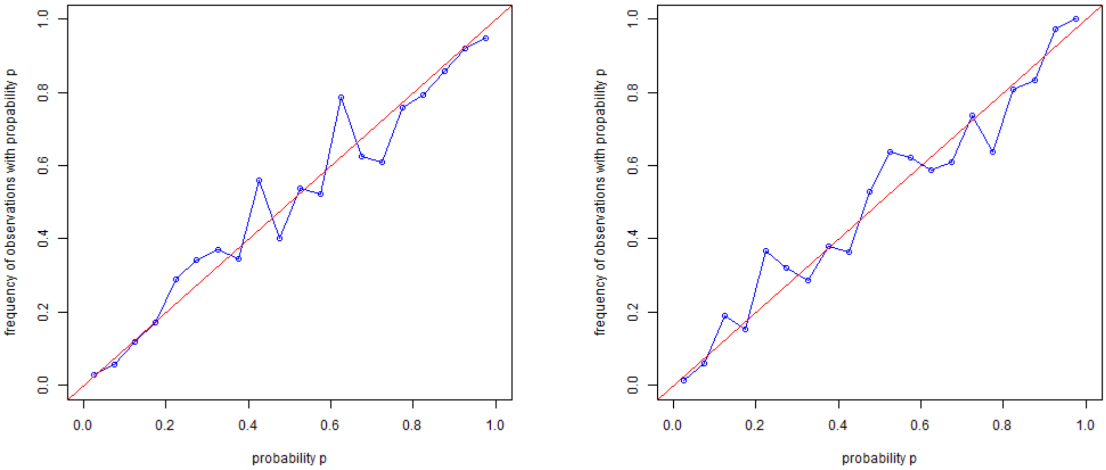


Figure 4.14.: Reliability diagrams of area precipitation probabilities for two selected test areas (Voronoi cells).

Ideally, the relative frequency of precipitation among all forecast periods with area probabilities in the subinterval U_k should again fall into U_k for $k = 1, \ldots, 20$. Thus, a high forecast quality is characterized by a reliability diagram being close to the curve $\{(x, y) \in [0, 1]^2 : x = y\}$.

In contrast to the score functions considered in Section 4.7.1, it is not possible to show reliability diagrams for all Voronoi cells here. In Figure 4.14, diagrams for two sample Voronoi cells are depicted, which show a high correspondence of area probabilities and precipitation indicators. It is clearly shown for these examples that precipitation frequently occurs if computed precipitation probabilities are high and vice versa. Similar results were obtained for most other Voronoi cells, too. A possibility to assess performance of area probabilities for all test areas is to compute a reliability diagram according to (4.20) based on area probabilities and precipitation indicators of all Voronoi cells which intersect the restricted window W_{in} and all forecast periods simultaneously, see Figure 4.15. Again, we observe a great forecast performance with the exception that for probabilities between 0.7 and 0.9 the corresponding reliabilities are marginally too low. This indicates that those probabilities are forecasted slightly too often and should be reduced, which corresponds well with the average bias of 1 % in area probabilities. Altogether, the forecast verification performed in this section shows impressively that area precipitation probabilities computed using the proposed germ-grain model of precipitation cells correspond well with radar observations.

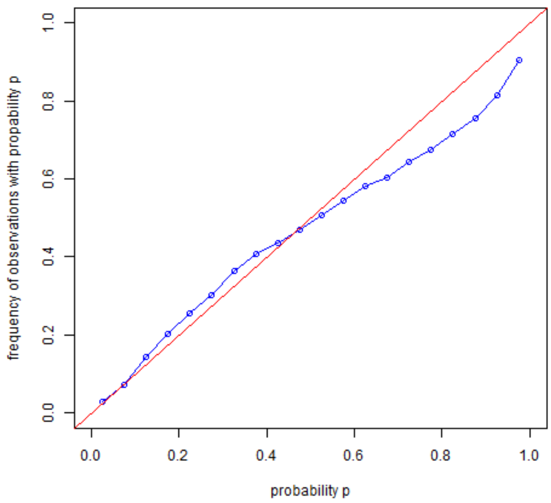


Figure 4.15.: Reliability diagram computed based on area precipitation probabilities of all Voronoi cells intersecting W_{in}.

5. Spatial stochastic modeling of precipitation amounts

In Chapter 4, a stochastic model for precipitation cells is introduced with the purpose of computing area probabilities for the occurrence of precipitation. To allow for a verification of obtained area forecasts using rain gauge adjusted radar measurements, we consider sequences of point probabilities for the occurrence of precipitation of more than 0.1 mm as data input since this is the smallest positive precipitation amount provided by the RADOLAN system. Accordingly, obtained area probabilities can also be understood as probabilities for the occurrence of precipitation of more than 0.1 mm somewhere inside the considered areas. Although there are a couple of applications where such area probabilities play an important role, meteorologists are rather interested in the probability of precipitation of more than u mm somewhere inside an area for an arbitrary threshold $u > 0$, see Section 1.1. For example, this is of particular relevance for the issuing of weather warnings, where extreme precipitation amounts need to be forecasted as precisely as possible. An intuitive approach to derive area probabilities for an arbitrary threshold $u > 0$ would be the following generalization of the models and methods introduced in Chapter 4. At first, the occurrence of precipitation is considered to be identical with the occurrence of precipitation of more than u mm, i.e., smaller precipitation amounts are ignored (this is done for $u = 0.1$ mm in Chapter 4). Then, a spatial stochastic model for the union set of precipitation cells (only representing precipitation of more than u mm) is introduced as described in Section 4.4 and model characteristics are computed based on point probabilities for the occurrence of precipitation of more than u mm in a similar way as proposed in Section 4.5. Finally, the generalized model can be used to compute area probabilities for the occurrence of precipitation of more than u mm, compare to Section 4.6. However, it seems clear that if the threshold u is significantly larger than 0.1 mm, then there is a high probability that realizations of the random radius R of precipitation cells are considerably lower than in the case $u = 0.1$ mm. In particular, if extreme precipitation events are forecasted, e.g., if $u \geq 5$ mm, typical 'precipitation cells' are expected to be clearly smaller than the sizes of almost all Voronoi cells that correspond to the system of locations with available point data, which causes the proposed model to be significantly less efficient. To account for this problem, a much higher number of locations with available point probabilities is needed, which leads to a drastic increase in required computation time. Another disadvantage of this approach is that model characteristics need to be determined separately for all considered thresholds, which is highly inefficient, making it inappropriate for application in operational weather prediction. Thus, we do not further pursue this idea in the following.

In the present chapter, an alternative approach to the estimation of area probabilities for precipitation exceeding an arbitrary threshold $u > 0$ is introduced. We propose an extension of the model for precipitation cells discussed in Chapter 4 by adding a spatial stochastic model for precipitation amounts. Clearly, a modeling of precipitation amounts based solely on point probabilities for the occurrence of precipitation of more than 0.1 mm is not possible, which is why a more comprehensive data basis is needed, see Section 5.1. In Section 5.2, the construction of the above-mentioned spatial stochastic model of precipitation amounts is described and the model is embedded in the mathematical framework introduced in Chapter 4. The fitting of (time-dependent) model characteristics based on available point forecast data is discussed in Sections 5.3 and 5.4. As no direct computation formula is available, Section 5.5 provides a Monte Carlo approach to the estimation of area probabilities of precipitation exceeding arbitrary thresholds. Finally, Section 5.6 describes a verification of obtained area probabilities using adjusted radar data. In this context, we also explain how remaining (not time-dependent) model parameters are fitted. The results presented in this chapter have been incorporated in Kriesche et al. (2015b) (in a modified mathematical framework) and Kriesche et al. (2017c). While the model for the occurrence of precipitation introduced in Chapter 4 can reasonably be generalized to the representation of other weather events such as wind gusts, one has to be much more careful when aiming to do the same for the presented model of precipitation amounts as, e.g., intensities of wind gusts (which are an analogue to precipitation amounts) could have an entirely different spatial structure.

5.1. Description of data

In order to illustrate the proposed methodology for the estimation of area probabilities for precipitation amounts exceeding arbitrary thresholds, we use a similar data basis as in Chapter 4. We again consider the system of 503 German and Luxembourgian weather stations illustrated in Figure 4.1, whose locations are projected into a rectangle $[0, 1500] \times [0, 1875]$ of pixel coordinates. We recall that distances between locations inside this window are computed based on the two-dimensional Euclidean distance (which only produces a negligible error) and that any such pixel distance $d > 0$ approximately corresponds to $0.5 \cdot d$ km. Moreover, we again consider a time frame covering the period from June 1 until July 31 and November 1 until December 31 in the year 2012. At each day, forecasts are available for the 7 one-hour forecast periods 2-3 UTC, 5-6 UTC, $\ldots$, 20-21 UTC resulting in a total of 854 forecast periods. Using the MOSMIX system of DWD for post-processing, see Section 2.3.2, point probabilities are computed for each weather station and forecast period describing the chance of precipitation of more than u mm for a sequence of thresholds $u \in \mathbb{T} = \{0, 0.1, 0.2, 0.3, 0.5, 0.7, 1, 2, 3, 5, 10, 15\}$ in mm. In particular, for $u = 0$ the probability for precipitation of any amount at the considered weather station is given. Unfortunately, these point forecast data are not yet suitable for the purpose of model calibration. On the one hand, point probabilities for a fixed weather station and forecast period are expected to be monotonically decreasing with increasing threshold $u \in \mathbb{T}$. By using the MOS approach in the post-processing step, however, the probabilities of each threshold are computed separately, which does not guarantee monotonicity due to statistically independent noise. In a few cases in our data, it is possible that, e.g., the probability of precipitation of more than 0.3 mm at a given weather station is slightly higher than the probability of precipitation of more than 0.2 mm. On the other hand, we will see that the model of precipitation amounts proposed in Section 5.2 requires expectations and variances of point precipitation amounts to be available for model fitting, which are, however, not directly included in the data. To overcome both problems, an additional modeling approach for precipitation amounts at the locations of weather stations is introduced in Section 5.3. Finally, for the purpose of forecast verification the rain gauge adjusted radar data described in Section 4.1 are available, see the sample data for two selected forecast periods shown in Figure 4.3.

5.2. Combined modeling of precipitation cells and precipitation amounts

The derivation of area probabilities for the occurrence of precipitation amounts exceeding arbitrary thresholds based solely on a model of precipitation cells does not seem to be possible. Therefore, an extension of the model presented in Section 4 is proposed, which adds precipitation amounts to modeled precipitation cells. We recall the most important notation, with the difference that in this chapter the occurrence of precipitation is no longer identical to the occurrence of precipitation of more than 0.1 mm, which enables the consideration of smaller positive precipitation amounts. Let T be a fixed one-hour forecast period and let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space introduced in Section 4.2, where the sample space Ω contains all possible weather scenarios in forecast period T and the corresponding forecasts provided by DWD, $\mathcal{F}$ is a suitable σ-algebra of subsets of Ω, and $\mathbb{P}$ is some probability measure on $(\Omega, \mathcal{F})$. By $E : \Omega \to \mathbb{S}$ we denote the random error occurring in the weather forecast models of DWD and $\sigma(E) \subset \mathcal{F}$ is the sub-σ-algebra of events generated by E. We consider the random field $\{P(t), t \in W\}$ defined in a compact and convex domain $W \subset \mathbb{R}^2$, with $P(t) : \Omega \to [0, 1]$

being the random point probability for the occurrence of precipitation of any amount at location $t \in W$ in the one-hour forecast period T (note the difference to Section 4.3, where $P(t)$ describes the probability of precipitation of more than 0.1 mm). Furthermore, $P(t)$ is assumed to be $\sigma(E)$-measurable for all $t \in W$. By $W_{in} \subset W$ we denote a compact subset of W containing locations with a certain distance to the boundaries of Germany, which is defined in Section 4.5.2. We suppose that point forecasts are provided by DWD for a system of locations $s_1, \ldots, s_n \in W$ with $n \in \mathbb{N}$ and that available data include a sequence $p^{(0)}(s_1) = \mathbb{E}\left(P(s_1) \mid E = e\right), \ldots, p^{(0)}(s_n) = \mathbb{E}\left(P(s_n) \mid E = e\right)$ of point probabilities for the occurrence of precipitation of any amount at $s_1, \ldots, s_n$, which are computed based on a particular realization e of E. By $\{V(s_1), \ldots, V(s_n)\}$ we denote the Voronoi tessellation of $s_1, \ldots, s_n$ introduced in (4.2). In the example of application considered to illustrate the modeling approach presented in this chapter, we again have that $W = [0, 1500] \times [0, 1875]$ is a rectangle of projected pixel coordinates comprising Germany and Luxembourg, the sequence $s_1, \ldots, s_{503}$ identifies the $n = 503$ weather stations depicted in Figure 4.1, the point probabilities $p^{(0)}(s_1), \ldots, p^{(0)}(s_{503})$ are included in the data described in Section 5.1, and e is the particular error that occurs when computing these data. The union set of precipitation cells (describing precipitation of any amount) is represented by the random germ-grain model $M : \Omega \to \mathsf{C}$ specified in (4.3), which is characterized by the $\sigma(E)$-measurable random local intensities $A_1, \ldots, A_n : \Omega \to [0, \infty)$ of the underlying Cox process $\{X_i, i = 1, \ldots, Z\}$ of cell centers and the $\sigma(E)$-measurable random cell radius $R : \Omega \to (0, \infty)$, see Section 4.4.

Based on the random union set M of precipitation cells a spatial stochastic model for the representation of precipitation amounts is proposed. For that purpose, we introduce the random field $\{\Gamma(t), t \in W\}$, where $\Gamma(t) : \Omega \to [0, \infty)$ describes the random amount of precipitation at each location $t \in W$ during the forecast period T. We expect that precipitation cells and precipitation amounts cannot be considered to be independent of one another, which is indicated by the results of a statistical test performed in Koubek et al. (2016). We suggest to represent $\{\Gamma(t), t \in W\}$ as a random shot-noise field since this class of random fields has been successfully applied in the literature for the modeling of precipitation amounts before, see, e.g., Rodriguez-Iturbe et al. (1986). At first, a radially symmetric response function $\kappa_p(\cdot, X_i, R)$ is assigned to each precipitation cell $b(X_i, R)$ for $i = 1, \ldots, Z$, where we choose $\kappa_p : \mathbb{R}^2 \times \mathbb{R}^2 \times (0, \infty) \to [0, \infty)$ as

$$\kappa_p(t, x, r) = \left(1 - \frac{\|t - x\|_2^2}{r^2}\right)^p \mathbb{1}_{b(t,r)}(x), \quad t, x \in \mathbb{R}^2, r > 0, \tag{5.1}$$

with a certain shape parameter $p > 0$. This choice comprises a variety of possible response functions. For example, if $r = 1$ and $t = o$, we obtain the surface of the upper half of the 3-dimensional unit ball (for $p = 0.5$), a scaled version of the 2-dimensional radially symmetric Epanechnikov kernel (for $p = 1$), a scaled version of the 2-dimensional radially symmetric biweight kernel (for $p = 2$) or a scaled version of the 2-dimensional radially symmetric triweight kernel (for $p = 3$), compare to Section 3.4.1.

However, these response functions are not yet suitable to model precipitation fields generated by single precipitation cells as the distribution of precipitation amounts is expected to vary across the domain W for most forecast periods. As an example consider the periods June 16, 2012, 20-21 UTC and December 9, 2012, 11-12 UTC, both of which have higher expected precipitation amounts in northern and central Germany than in the south, see Figure 5.4 later on. To account for spatially varying distributions of precipitation amounts, we suggest to multiply each response function $\kappa_p(\cdot, X_i, R)$ by a random location-dependent scaling variable. Since point forecast data are only available at a finite sequence $s_1, \ldots, s_n$ of locations in W, we make the simplifying assumption that all precipitation cells with centers in a given Voronoi cell $V(s_i)$ are multiplied by the same scaling variable for $i = 1, \ldots, n$. Thus, we introduce a sequence $C_1, \ldots, C_n : \Omega \to [0, \infty)$ of nonnegative, absolutely continuous random scaling variables, which correspond to the n Voronoi cells $V(s_1), \ldots, V(s_n)$. The variables $C_1, \ldots, C_n$ can clearly not be assumed to be $\sigma(E)$-measurable since precipitation amounts are still expected to be random if a particular realization of the weather forecast models of DWD is given. However, we assume that conditioned on E, the variables $C_1, \ldots, C_n$ are independent of each other and of the point process $\{X_i, i = 1, \ldots, Z\}$ of precipitation cell centers. For each $t \in W$, we interpret the value of the response function $\kappa_p(t, X_i, R)$ multiplied by the corresponding scaling variable as the random amount of precipitation generated by the i-th precipitation cell $b(X_i, R)$ at location t. The total amount of precipitation at $t \in W$ is obtained by summing up the individual precipitation amounts generated by all Z precipitation cells. Combining the modeling steps suggested above leads to the following representation of the random field $\{\Gamma(t), t \in W\}$ of precipitation amounts:

$$\Gamma(t) = \sum_{i=1}^{Z} \sum_{j=1}^{n} C_j \mathbb{1}_{V(s_j)}(X_i) \kappa_p(t, X_i, R), \quad t \in W. \tag{5.2}$$

The consecutive steps of modeling precipitation amounts according to (5.2) are exemplified in Figure 5.1.

5.3. Distribution of precipitation amounts at data locations

The random field $\{\Gamma(t), t \in W\}$ of precipitation amounts introduced in (5.2) is completely characterized by the random intensities $A_1, \ldots, A_n$ for the occurrence of precipitation cells, the random cell radius R, the random scaling variables $C_1, \ldots, C_n$, and the shape parameter p. In the following, let e be the particular realization of the random error E that occurs when computing the underlying data using the weather forecast models of DWD. Recall that the (conditional) intensities $a_1 =$

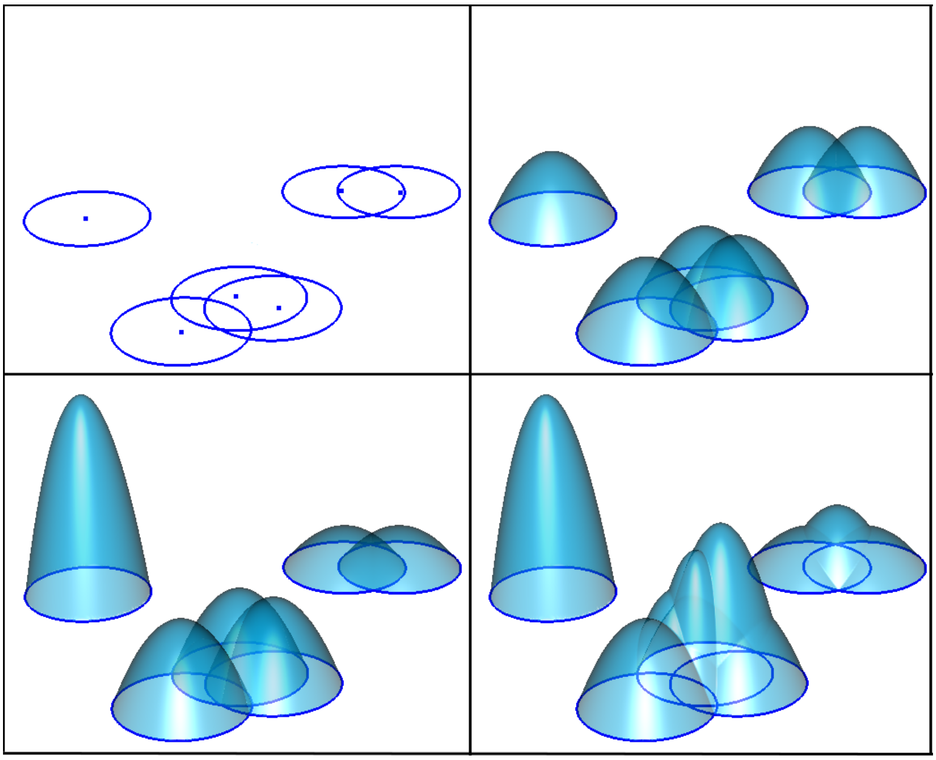


Figure 5.1.: Illustration of the proposed modeling approach for precipitation amounts: (i) modeling of precipitation cells using a germ-grain model with circular grains (top left), (ii) assigning a radially symmetric response function to each cell (top right), (iii) multiplying response functions with random scaling variables (bottom left), (iv) summing scaled response functions to obtain precipitation amounts (bottom right).

$\mathbb{E}(A_1 \,|\, E = e), \ldots, a_n = \mathbb{E}(A_n \,|\, E = e)$ and the cell radius $r = \mathbb{E}(R \,|\, E = e)$ can be computed according to (4.7) and (4.12) based on the corresponding point probabilities $p^{(0)}(s_1) = \mathbb{E}(P(s_1) \,|\, E = e), \ldots, p^{(0)}(s_n) = \mathbb{E}(P(s_n) \,|\, E = e)$. In order to use the model for the estimation of area probabilities, it remains to fit (conditional) distributions of the random scaling variables $C_1, \ldots, C_n$ (given $\{E = e\}$) and to choose a suitable shape parameter p. For that purpose, we first introduce the deterministic fields $\{\epsilon(t), t \in W\}$ and $\{v(t), t \in W\}$, where $\epsilon(t) = \mathbb{E}(\Gamma(t) \,|\, E = e) \in [0, \infty)$ and $v(t) = \mathrm{var}(\Gamma(t) \,|\, E = e) \in [0, \infty)$ denote the conditional expectation and variance of $\Gamma(t)$ given $\{E = e\}$ for all $t \in W$. We assume that point probabilities for the occurrence of precipitation of more than u mm in the considered forecast period T are included in the data provided by DWD for all locations $s_1, \ldots, s_n$ and thresholds $u \in \mathbb{T} = \{0, 0.1, 0.2, 0.3, 0.5, 0.7, 1, 2, 3, 5, 10, 15\}$. However, in general these data are not yet suitable for model fitting due to occasional inconsistencies in point probabilities and missing expectations and variances of point precipitation amounts, see, e.g., the sample data described in Section 5.1.

To overcome such potential problems, we suggest to fit a gamma distribution to the random point precipitation amount at each location $s \in \{s_1, \dots, s_n\}$ as proposed, e.g., in Wilks and Eggleston (1992) and Wilks (2011), Section 4.3.3. However, since the gamma distribution is an absolutely continuous distribution and $P(\Gamma(s) = 0 \mid E = e) > 0$ in general, it is an inappropriate choice for the distribution of $\Gamma(s)$ (given $\{E = e\}$). Instead, we assume that the gamma distribution can be used to model the (positive) precipitation amount $\Gamma(s)$ if precipitation occurs at s, i.e., given $\{\Gamma(s) > 0\}$. For that purpose, let $p^{(u)}(s) = P(\Gamma(s) > u \mid E = e)$ for each $u \in \mathbb{T}$ denote the (conditional) probability for the occurrence of precipitation of more than u mm (given $\{E = e\}$) at location s and $\tilde{p}^{(u)}(s) = P(\Gamma(s) > u \mid \Gamma(s) > 0, E = e)$ the conditional probability for the occurrence of precipitation of more than u mm given that precipitation of any (positive) amount occurs at s for each threshold $u \in \mathbb{T} \setminus \{0\}$. Note that this corresponds well to the definition of $p^{(0)}(s)$ as given in Section 5.2 since $\{\Gamma(s) > 0\} = \{s \in M\}$, compare to the modeling of point probabilities as coverage probabilities of M in (4.1) and the definition of $\{\Gamma(t), t \in W\}$ in (5.2). We assume that $p^{(0)}(s) > 0$ as otherwise $\Gamma(s) = 0$ a.s. and $p^{(u)}(s) = 0$ for all $u \in \mathbb{T} \setminus \{0\}$. By using that for any nonnegative random variable $X : \Omega \to [0, \infty)$ with $P(X > 0) > 0$ and any $u > 0$ it holds that $P(X > u \mid X > 0) = P(X > u)/P(X > 0)$, the conditional probabilities of precipitation exceeding u mm given that precipitation of any amount occurs at s can be determined based on $p^{(0)}(s)$ and the probabilities for the occurrence of precipitation of more than u mm from the available data for all $u \in \mathbb{T} \setminus \{0\}$. Based on the obtained conditional probabilities the parameters of a gamma distribution are fitted. This allows to analytically compute the sequence $\{\tilde{p}^{(u)}(s), u \in \mathbb{T} \setminus \{0\}\}$ of (conditional) probabilities according to the fitted gamma distribution. Furthermore, we also compute the conditional expectations $\tilde{\epsilon}(s) = \mathbb{E}\,(\Gamma(s) \mid \Gamma(s) > 0, E = e)$ and $\tilde{m}(s) = \mathbb{E}\,(\Gamma^2(s) \mid \Gamma(s) > 0, E = e)$ based on the obtained gamma distribution. Then, the point probabilities $\{p^{(u)}(s), u \in \mathbb{T} \setminus \{0\}\}$ can be recomputed easily using the identity

$$p^{(u)}(s) = p^{(0)}(s)\,\tilde{p}^{(u)}(s), \quad u \in \mathbb{T} \setminus \{0\}, \tag{5.3}$$

see above. This approach has the advantage that obtained point probabilities are now monotonically decreasing with increasing threshold u. Moreover, the fitted gamma distribution allows to easily compute the expectation $\epsilon(s)$ and variance $v(s)$ of $\Gamma(s)$ (conditioned on $\{E = e\}$). Using (5.3) and a representation formula for the moments of a nonnegative random variable based on its tail function, see Kallenberg (2002), Lemma 3.4, we get that

$$\tilde{\epsilon}(s) = \int_0^\infty \tilde{p}^{(u)}(s)\,\mathrm{d}u = \int_0^\infty \frac{p^{(u)}(s)}{p^{(0)}(s)}\,\mathrm{d}u = \frac{\epsilon(s)}{p^{(0)}(s)},$$

which allows to determine $\epsilon(s)$ according to

$$\epsilon(s) = p^{(0)}(s)\,\tilde{\epsilon}(s). \tag{5.4}$$

Similarly, it holds that

$$\tilde{m}(s) = 2\int_0^\infty \tilde{p}^{(u)}(s)\, u \,\mathrm{d}u = 2\int_0^\infty \frac{p^{(u)}(s)}{p^{(0)}(s)}\, u\,\mathrm{d}u = \frac{\mathbb{E}\left(\Gamma^2(s)\,|\,E=e\right)}{p^{(0)}(s)},$$

which is used to compute $v(s)$ by

$$v(s) = \mathbb{E}\left(\Gamma^2(s)\,|\,E=e\right) - \epsilon^2(s) = p^{(0)}(s)\,\tilde{m}(s) - \left(p^{(0)}(s)\,\tilde{\epsilon}(s)\right)^2. \qquad (5.5)$$

The suggested fitting procedure provides consistent (conditional) point probabilities for the occurrence of precipitation of more than u mm for thresholds $u \in \mathbb{T}$ as well as (conditional) expectations and variances of point precipitation amounts (given $\{E = e\}$) at all data locations $s_1, \ldots, s_n$. These revised data are now suitable to be used for the calibration of the proposed model for precipitation amounts introduced in (5.2), see Sections 5.4 and 5.6. To illustrate the results that are obtained in our example of application by fitting gamma distributions as described above, we present some sample data for two selected forecast periods. Figures 5.2 and 5.3 show point probabilities $p^{(u)}(s_1), \ldots, p^{(u)}(s_{503})$ for thresholds $u \in \{0, 0.2, 0.7, 2\}$ in mm, which are computed according to (5.3) based on the available data described in Section 5.1. In both examples, point probabilities decrease rapidly for increasing thresholds but for each single threshold there are generally higher probabilities of precipitation in forecast period December 9, 2012, 11-12 UTC than in June 16, 2012, 20-21 UTC. Even though the point probabilities shown in Figure 4.2 are not identical to $p^{(0.1)}(s_1), \ldots, p^{(0.1)}(s_{503})$ as they are directly derived from the weather prediction models of DWD without additionally fitting a gamma distribution, they correspond well to the illustrations given in this section. Obtained probabilities of precipitation exceeding 5, 10 or 15 mm are close to zero for almost all weather stations and forecast periods, which is why the occurrence of such precipitation amounts is considered to be an extreme event. In Figure 5.4, expected precipitation amounts $\epsilon(s_1), \ldots, \epsilon(s_{503})$ determined according to (5.4) are depicted for the same forecast periods. As anticipated, expected precipitation amounts seem strongly correlated to point probabilities showing larger values in regions of higher point probabilities (for all thresholds) and vice versa. Furthermore, we observe a generally good accordance to the corresponding radar data, compare to Figure 4.3. Finally, Figure 5.5 illustrates variances $v(s_1), \ldots, v(s_{503})$ computed according to (5.5) for the two considered sample forecast periods. For each fixed period we find a certain association between expectations and variances of point precipitation amounts as variances seem to be largest for locations with large expected precipitation amounts and vice versa. However, this does not seem to be valid when comparing different forecast periods. Although expected precipitation amounts are clearly larger in December 9, 2012, 11-12 UTC than in June 16, 2012, 20-21 UTC (in particular for northern Germany), we observe similarly large variances of precipitation amounts in both forecast periods.

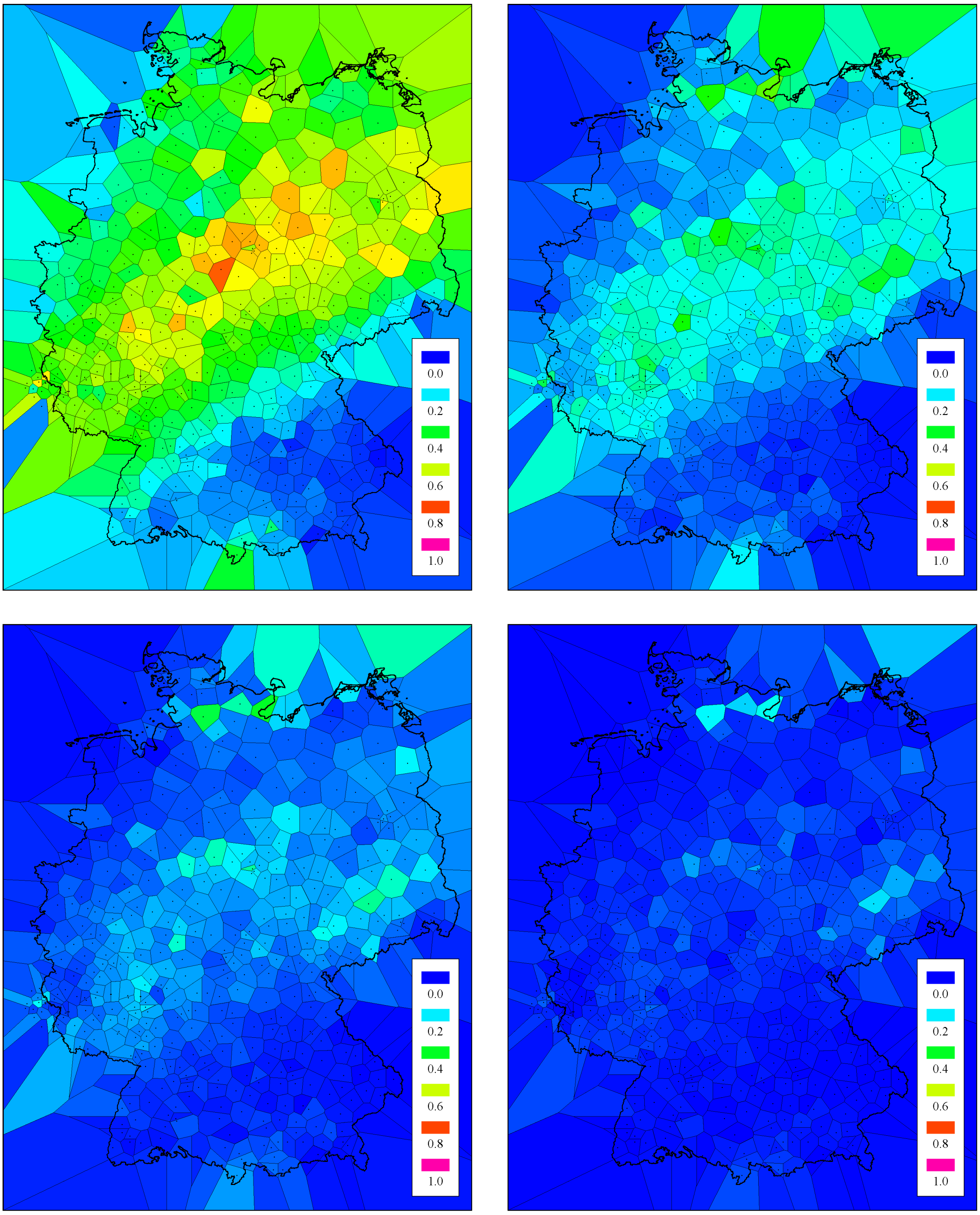


Figure 5.2.: Sample data for June 16, 2012, 20-21 UTC: the locations of the considered 503 weather stations and the corresponding Voronoi tessellation, where each Voronoi cell is colored according to the point probability for the occurrence of precipitation of more than 0 mm (top left), 0.2 mm (top right), 0.7 mm (bottom left), and 2 mm (bottom right) at the corresponding station. Point probabilities are obtained from fitted gamma distributions.

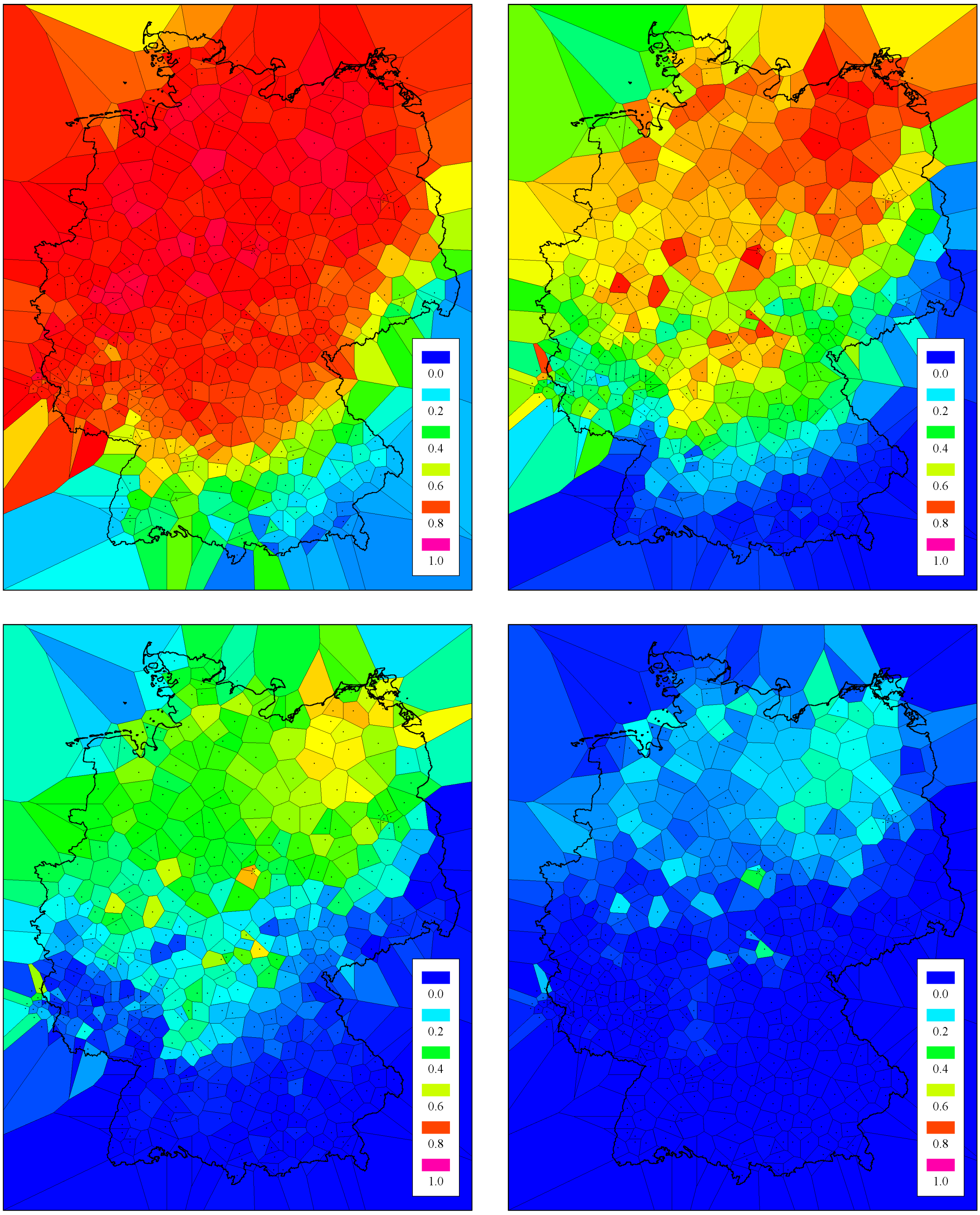
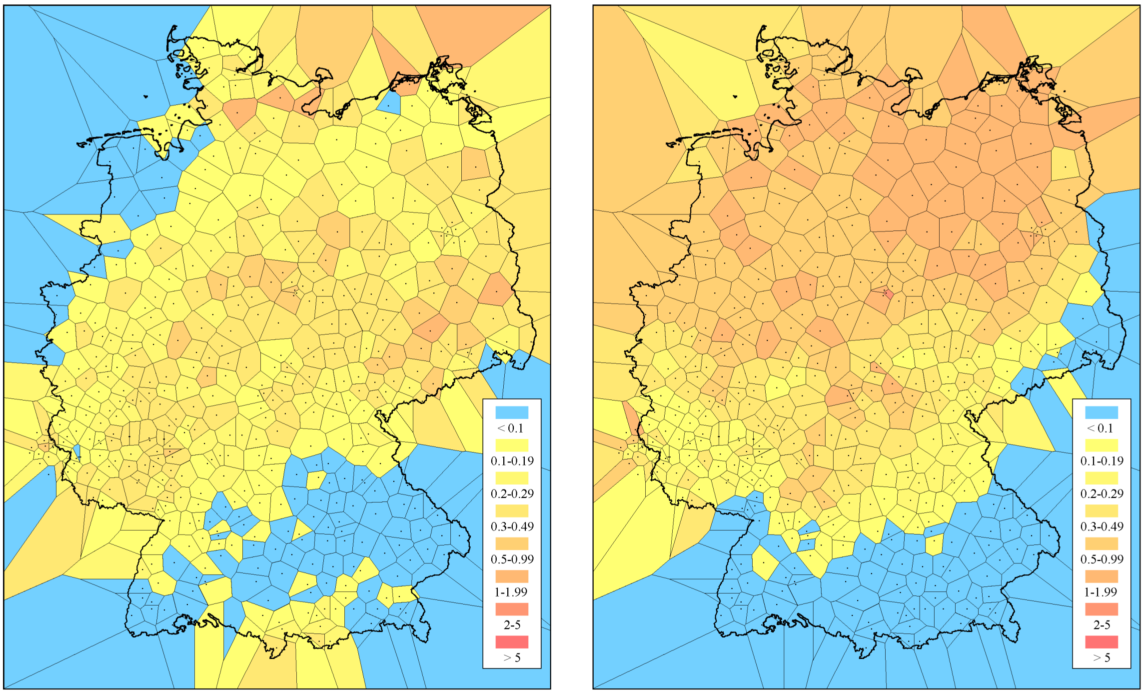


Figure 5.3.: Sample data for December 9, 2012, 11-12 UTC: the locations of the considered 503 weather stations and the corresponding Voronoi tessellation, where each Voronoi cell is colored according to the point probability for the occurrence of precipitation of more than 0 mm (top left), 0.2 mm (top right), 0.7 mm (bottom left), and 2 mm (bottom right) at the corresponding station. Point probabilities are obtained from fitted gamma distributions.



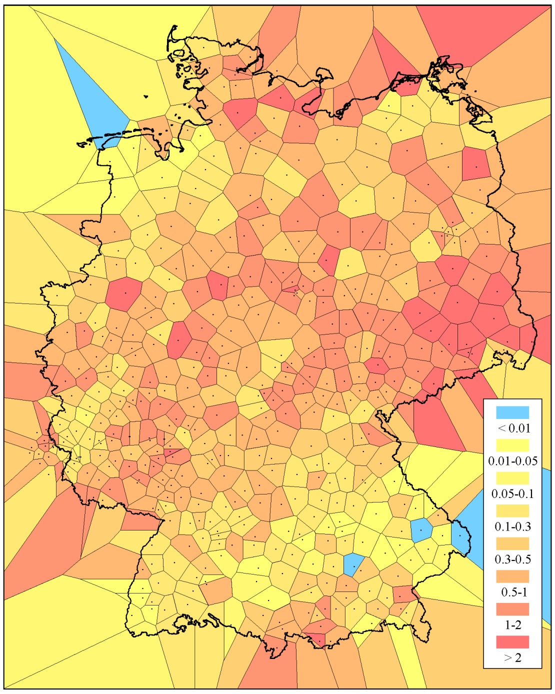
June 16, 2012, 20-21 UTC

December 9, 2012, 11-12 UTC

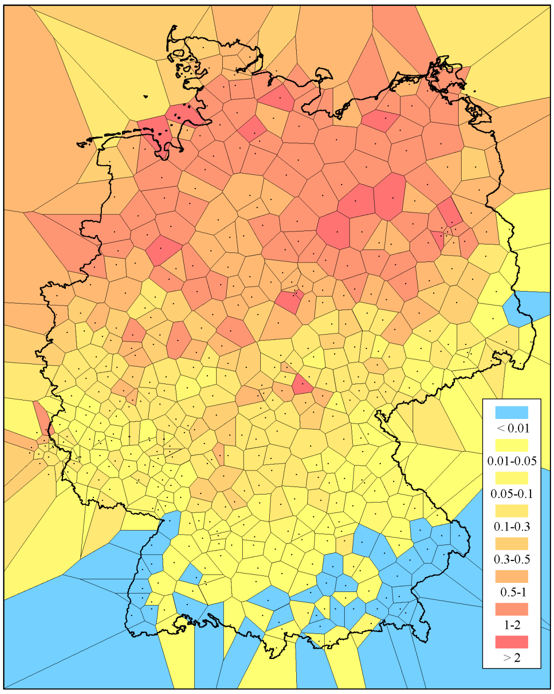
Figure 5.4.: Sample data for two selected forecast periods: the locations of the considered 503 weather stations and the corresponding Voronoi tessellation, where each Voronoi cell is colored according to the expected precipitation amount at the corresponding station. Expected precipitation amounts are obtained from fitted gamma distributions.

5.4. Fitting conditional distributions of random scaling variables

We now introduce an approach to fitting (conditional) distributions of the random scaling variables $C_1, \ldots, C_n$ (given $\{E = e\}$) based on the (conditional) expectations $\epsilon(s_1), \ldots, \epsilon(s_n)$ and variances $v(s_1), \ldots, v(s_n)$ of precipitation amounts at $s_1, \ldots, s_n$ that were derived according to (5.4) and (5.5) using fitted gamma distributions. The goal of the method is to choose the distributions of $C_1, \ldots, C_n$ in such a way that expectations and variances of precipitation amounts provided by the calibrated model stated in (5.2) are as close as possible to the derived data $\epsilon(s_1), \ldots, \epsilon(s_n)$ and $v(s_1), \ldots, v(s_n)$. We will always assume that the shape parameter p of the response function κ_p defined in (5.1) is fixed. A recommendation on how to choose p in practice is given later on in Section 5.6.



June 16, 2012, 20-21 UTC



December 9, 2012, 11-12 UTC

Figure 5.5.: Sample data for two selected forecast periods: the locations of the considered 503 weather stations and the corresponding Voronoi tessellation, where each Voronoi cell is colored according to the variance of the precipitation amount at the corresponding station. Variances of precipitation amounts are obtained from fitted gamma distributions.

5.4.1. Conditional expectations and variances of scaling variables

At first, we describe a statistical method to compute (conditional) expectations and variances of $C_1, \ldots, C_n$. For that purpose, recall that conditioned on $\{E = e\}$ the point process $\{X_i, i = 1, \ldots, Z\}$ of precipitation cell centers is a Poisson point process with intensity function $\{\lambda(t), t \in W\}$, where $\lambda(t) = \mathbb{E}\,(\Lambda(t) \,|\, E = e)$ for all $t \in W$ with $\{\Lambda(t), t \in W\}$ being defined as in Section 4.4. Additionally, the scaling variables $C_1, \ldots, C_n$ are assumed to be conditionally independent of each other and of $\{X_i, i = 1, \ldots, Z\}$ given $\{E = e\}$. To improve the readability of derived representation formulas, we introduce the following simplified notation. For each $j \in \{1, \ldots, n\}$, let $c_j = \mathbb{E}\,(C_j \,|\, E = e)$ and $\tilde{c}_j = \mathrm{var}\,(C_j \,|\, E = e)$ denote the conditional expectation and variance of C_j conditioned on $\{E = e\}$. Furthermore, we introduce the functions $f_j : W \times W \to [0, \infty)$, $J_j : W \to [0, \infty)$ and $\tilde{J}_j : W \to [0, \infty)$ that are defined by

$$f_j(t, x) = \mathbb{1}_{V(s_j)}(x)\, \kappa_p(t, x, r), \quad t, x \in W,$$

$$J_j(t) = \int_{V(s_j)\cap b(t,r)} \left(1 - \frac{\|t-x\|_2^2}{r^2}\right)^p \mathrm{d}x = \int_W f_j(t,x)\,\mathrm{d}x, \quad t \in W,$$

and

$$\tilde{J}_j(t) = \int_{V(s_j)\cap b(t,r)} \left(1 - \frac{\|t-x\|_2^2}{r^2}\right)^{2p} \mathrm{d}x = \int_W f_j^2(t,x)\,\mathrm{d}x, \quad t \in W.$$

We start by deriving a representation formula of the conditional expectation $\epsilon(t) = \mathbb{E}\left(\Gamma(t) \,|\, E = e\right)$ for $t \in W$. By applying the Campbell theorem for random point processes, see Theorem 3.3.2, with $f(x) = f_j(t,x)$ and by using that $\mathbb{1}_{V(s_j)}(x)\mathbb{1}_{V(s_k)}(x) = 0$ if $j \neq k$ we get that

$$\begin{aligned}
\epsilon(t) &= \mathbb{E}\left(\sum_{i=1}^{Z}\sum_{j=1}^{n} C_j \mathbb{1}_{V(s_j)}(X_i)\kappa_p(t, X_i, R) \,\Big|\, E = e\right) \\
&= \sum_{j=1}^{n} c_j\, \mathbb{E}\left(\sum_{i=1}^{Z} f_j(t, X_i) \,\Big|\, E = e\right) \\
&= \sum_{j=1}^{n} c_j \int_W f_j(t,x) \sum_{k=1}^{n} a_k \mathbb{1}_{V(s_k)}(x)\,\mathrm{d}x \\
&= \sum_{j=1}^{n} c_j\, a_j\, J_j(t), \quad t \in W.
\end{aligned} \tag{5.6}$$

In particular, (5.6) should be satisfied for $t = s_1, \ldots, s_n$, which results in a system of n linear equations with unknown variables $c_1, \ldots, c_n$ (recall that $a_1, \ldots, a_n$ and r were already computed in the context of calibrating the germ-grain model M of precipitation cells and p is assumed to be known). In general, this system of equations cannot be solved exactly under the constraint that $c_1, \ldots, c_n \geq 0$ (as $C_1, \ldots, C_n \geq 0$ a.s.). Thus, we suggest to compute $c_1, \ldots, c_n$ in a nonnegative least-squares sense, i.e., by

$$(c_1, \ldots, c_n) = \underset{c'_1,\ldots,c'_n \geq 0}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(\epsilon(s_i) - \sum_{j=1}^{n} c'_j\, a_j J_j(s_i)\right)^2 \right\}, \tag{5.7}$$

where again the algorithm presented in Lawson and Hanson (1974), Chapter 23 can be used.

Next, we derive a representation formula of the conditional variance $v(t) = \operatorname{var}\left(\Gamma(t) \,|\, E = e\right)$ for $t \in W$, which is a more complicated task. As a first result we get that

$$\mathbb{E}\left(\Gamma^2(t) \,\Big|\, E = e\right) = \mathbb{E}\left(\left(\sum_{i=1}^{Z}\sum_{j=1}^{n} C_j \mathbb{1}_{V(s_j)}(X_i)\kappa_p(t, X_i, R)\right)^2 \,\Big|\, E = e\right)$$

$$
\begin{aligned}
&= \mathbb{E}\left(\sum_{j=1}^{n}\sum_{k=1}^{n} C_j C_k \sum_{i=1}^{Z} \mathbb{1}_{V(s_j)}(X_i)\kappa_p(t, X_i, R) \sum_{l=1}^{Z} \mathbb{1}_{V(s_k)}(X_l)\kappa_p(t, X_l, R) \,\Big|\, E = e\right)\\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} \mathbb{E}\left(C_j C_k \,|\, E = e\right)\, \mathbb{E}\left(\sum_{i=1}^{Z} f_j(t, X_i) \sum_{l=1}^{Z} f_k(t, X_l) \,\Big|\, E = e\right)\\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} \mathbb{E}\left(C_j C_k \,|\, E = e\right) \cdot\\
&\qquad \cdot \left(\int_W f_j(t,x)\, a_j \,\mathrm{d}x \int_W f_k(t,x)\, a_k \,\mathrm{d}x + \int_W f_j(t,x)\, f_k(t,x) \sum_{m=1}^{n} a_m \mathbb{1}_{V(s_m)}(x)\,\mathrm{d}x\right)\\
&= \sum_{j=1}^{n} \mathbb{E}\left(C_j^2 \,|\, E = e\right)\left(a_j^2\, J_j^2(t) + a_j\, \tilde{J}_j(t)\right) + \sum_{j=1}^{n}\sum_{\substack{k=1\\ k\neq j}}^{n} c_j\, c_k\, a_j\, a_k\, J_j(t)\, J_k(t), \quad t \in W,
\end{aligned}
$$

where we use the property of Poisson processes shown in Theorem 3.3.5 with $f(x) = f_j(t,x)$ and $g(x) = f_k(t,x)$ as well as that $f_j(t,x)f_k(t,x) = 0$ for all $t, x \in W$ if $j \neq k$. Furthermore, based on the representation formula of $\{\epsilon(t), t \in W\}$ derived in (5.6) it holds that

$$
\begin{aligned}
\left(\mathbb{E}\left(\Gamma(t) \,|\, E = e\right)\right)^2 &= \left(\sum_{j=1}^{n} c_j\, a_j\, J_j(t)\right)^2\\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} c_j\, c_k\, a_j\, a_k\, J_j(t)\, J_k(t)\\
&= \sum_{j=1}^{n} c_j^2\, a_j^2\, J_j^2(t) + \sum_{j=1}^{n}\sum_{\substack{k=1\\ k\neq j}}^{n} c_j\, c_k\, a_j\, a_k\, J_j(t)\, J_k(t), \quad t \in W.
\end{aligned}
$$

Finally, combining both results yields that

$$
\begin{aligned}
v(t) &= \mathbb{E}\left(\Gamma^2(t) \,\Big|\, E = e\right) - \left(\mathbb{E}\left(\Gamma(t) \,|\, E = e\right)\right)^2\\
&= \sum_{j=1}^{n} \mathbb{E}\left(C_j^2 \,|\, E = e\right)\left(a_j^2\, J_j^2(t) + a_j\, \tilde{J}_j(t)\right) - c_j^2\, a_j^2\, J_j^2(t)\\
&= \sum_{j=1}^{n} \mathbb{E}\left(C_j^2 \,|\, E = e\right)\left(a_j^2\, J_j^2(t) + a_j\, \tilde{J}_j(t)\right) - c_j^2\, \left(a_j^2\, J_j^2(t) + a_j\, \tilde{J}_j(t)\right) + c_j^2 a_j \tilde{J}_j(t)\\
&= \sum_{j=1}^{n} \tilde{c}_j\left(a_j \tilde{J}_j(t) + a_j^2 J_j^2(t)\right) + \sum_{j=1}^{n} c_j^2 a_j \tilde{J}_j(t), \quad t \in W. \qquad (5.8)
\end{aligned}
$$

Putting $t = s_1, \ldots, s_n$ in (5.8) again results in a system of n linear equations with unknown variables $\tilde{c}_1, \ldots, \tilde{c}_n$. Due to the constraint $\tilde{c}_1, \ldots, \tilde{c}_n \geq 0$ we solve the system of equations in a nonnegative least-squares sense, too, i.e.,

$$(\tilde{c}_1, \dots, \tilde{c}_n) = \underset{c_1', \dots, c_n' \geq 0}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(v(s_i) - \sum_{j=1}^{n} c_j^2 a_j \tilde{J}_j(s_i) \right.\right.$$
$$\left.\left. - \sum_{j=1}^{n} c_j' \left(a_j \tilde{J}_j(s_i) + a_j^2 J_j^2(s_i) \right) \right)^2 \right\}. \quad (5.9)$$

5.4.2. Suitable distributions for scaling variables

After having computed the (conditional) expectations $c_1, \dots, c_n$ and variances $\tilde{c}_1, \dots, \tilde{c}_n$ of the local scaling variables $C_1, \dots, C_n$ (given $\{E = e\}$), a two-parameter distribution can be fitted to each C_i for $i = 1, \dots, n$ using the method of moments. The following absolutely continuous distributions seem to be the most suitable ones since they are defined on the nonnegative real line, have finite second moments, and their parameters can be represented as closed functions of expectation and variance, which is required for applying the method of moments (two of the considered distributions have finite second moments for certain parameter configurations only).

Let $C \in \{C_1, \dots, C_n\}$ be a fixed random scaling variable and let $f : \mathbb{R} \to [0, \infty)$ denote the PDF of C.

1. The scaling variable C has a *beta prime distribution* with parameters $\alpha, \beta > 0$ if

   $$f(x) = \frac{x^{\alpha-1}(1+x)^{-\alpha-\beta}}{B(\alpha, \beta)} \mathbb{1}_{(0,\infty)}(x), \quad x \in \mathbb{R},$$

   where $B : (0, \infty) \times (0, \infty) \to (0, \infty)$ is the beta function defined as

   $$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt, \quad x, y > 0.$$

   Then, the expectation and variance of C are given by

   $$\mathbb{E}\, C = \frac{\alpha}{\beta - 1} \ \text{ for } \beta > 1 \quad \text{and} \quad \operatorname{var} C = \frac{\alpha(\alpha + \beta - 1)}{(\beta - 2)(\beta - 1)^2} \ \text{ for } \beta > 2.$$

   Accordingly, the parameters α and β can be computed based on $\mathbb{E}\, C$ and $\operatorname{var} C$ by

   $$\alpha = \frac{(\mathbb{E}\, C)^2(\mathbb{E}\, C + 1)}{\operatorname{var} C} + \mathbb{E}\, C \quad \text{and} \quad \beta = \frac{\mathbb{E}\, C(\mathbb{E}\, C + 1)}{\operatorname{var} C} + 2.$$

2. The scaling variable C has a *gamma distribution* with parameters $\alpha, \beta > 0$ if
$$f(x) = \frac{\beta^\alpha}{\Gamma'(\alpha)}\, x^{\alpha-1} e^{-\beta x}\, \mathbb{1}_{(0,\infty)}(x), \quad x \in \mathbb{R},$$
where $\Gamma' : (0,\infty) \to (0,\infty)$ is the gamma function defined as
$$\Gamma'(x) = \int_0^\infty t^{x-1} e^{-t}\, dt, \quad x > 0.$$
Then, the expectation and variance of C are given by
$$\mathbb{E}\, C = \frac{\alpha}{\beta} \quad \text{and} \quad \operatorname{var} C = \frac{\alpha}{\beta^2}.$$
Accordingly, the parameters α and β can be computed based on $\mathbb{E}\, C$ and $\operatorname{var} C$ by
$$\alpha = \frac{(\mathbb{E}\, C)^2}{\operatorname{var} C} \quad \text{and} \quad \beta = \frac{\mathbb{E}\, C}{\operatorname{var} C}.$$

3. The scaling variable C has an *inverse gamma distribution* with parameters $\alpha, \beta > 0$ if
$$f(x) = \frac{\beta^\alpha}{\Gamma'(\alpha)}\, x^{-\alpha-1} e^{-\frac{\beta}{x}}\, \mathbb{1}_{(0,\infty)}(x), \quad x \in \mathbb{R}.$$
Then, the expectation and variance of C are given by
$$\mathbb{E}\, C = \frac{\beta}{\alpha - 1} \ \text{ for } \alpha > 1 \quad \text{and} \quad \operatorname{var} C = \frac{\beta^2}{(\alpha-2)(\alpha-1)^2} \ \text{ for } \alpha > 2.$$
Accordingly, the parameters α and β can be computed based on $\mathbb{E}\, C$ and $\operatorname{var} C$ by
$$\alpha = \frac{(\mathbb{E}\, C)^2}{\operatorname{var} C} + 2 \quad \text{and} \quad \beta = \frac{(\mathbb{E}\, C)^3}{\operatorname{var} C} + \mathbb{E}\, C.$$

4. The scaling variable C has an *inverse normal distribution* (or *Wald distribution*) with parameters $\mu, \lambda > 0$ if
$$f(x) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\} \mathbb{1}_{(0,\infty)}(x), \quad x \in \mathbb{R}.$$
Then, the expectation and variance of C are given by
$$\mathbb{E}\, C = \mu \quad \text{and} \quad \operatorname{var} C = \frac{\mu^3}{\lambda}.$$
Accordingly, the parameters μ and λ can be computed based on $\mathbb{E}\, C$ and $\operatorname{var} C$ by
$$\mu = \mathbb{E}\, C \quad \text{and} \quad \lambda = \frac{(\mathbb{E}\, C)^3}{\operatorname{var} C}.$$

5. The scaling variable C has a *log-normal distribution* with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ if

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\} \mathbb{1}_{(0,\infty)}(x), \quad x \in \mathbb{R}.$$

Then, the expectation and variance of C are given by

$$\mathbb{E}\, C = e^{\mu + \frac{1}{2}\sigma^2} \quad \text{and} \quad \operatorname{var} C = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right).$$

Accordingly, the parameters μ and σ^2 can be computed based on $\mathbb{E}\, C$ and $\operatorname{var} C$ by

$$\mu = 2\log(\mathbb{E}\, C) - \frac{1}{2}\log(\operatorname{var} C + (\mathbb{E}\, C)^2) \quad \text{and} \quad \sigma^2 = \log\left(\frac{\operatorname{var} C}{(\mathbb{E}\, C)^2} + 1\right).$$

Finally, all (conditional) characteristics of the random field $\{\Gamma(t), t \in W\}$ of precipitation amounts that are assumed to depend on the current forecast period T have been determined: the local intensities $a_1, \ldots, a_n$ for the occurrence of precipitation cells, the cell radius r, and the (conditional) distributions of the local scaling variables $C_1, \ldots, C_n$. Note that the type of these distributions (one of the five introduced above) as well as the shape parameter p of the response function κ_p are considered to be fixed for all forecast periods. A recommendation on how to choose these remaining model configurations in practice is given in Section 5.6.

5.5. Model-based estimation of area probabilities

The proposed modeling approach for the representation of precipitation amounts in terms of the random field $\{\Gamma(t), t \in W\}$ introduced in (5.2) can be used for the derivation of area probabilities for the occurrence of precipitation exceeding arbitrary thresholds. According to our model, there is precipitation of more than $u \geq 0$ mm somewhere inside a Borel set $B \in \mathcal{B}(W)$ if $\Gamma(t) > u$ for at least one location $t \in B$. Thus, the $\sigma(E)$-measurable random area probability $\Pi^{(u)}(B) : \Omega \to [0,1]$ for the occurrence of precipitation of more than u mm in B is modeled as

$$\Pi^{(u)}(B) = \mathbb{P}(\max\{\Gamma(t), t \in B\} > u \mid E).$$

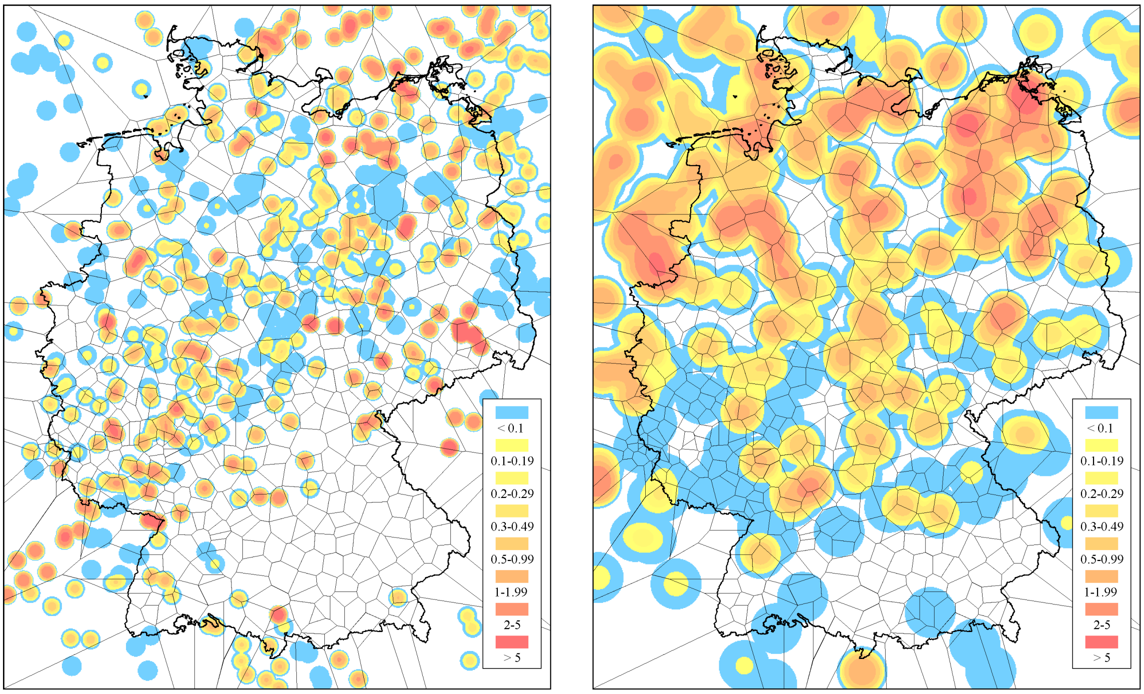
However, an analytical formula for $\Pi^{(u)}(B)$ (and also for point probabilities, which are obtained when B is chosen to be a singleton) can only be given for the occurrence of precipitation (i.e., for $u = 0$ mm), see (4.13). Therefore, we suggest to estimate area probabilities based on repeated Monte Carlo simulation. In applications (with a fixed realization e of E), we first determine the intensities $a_1, \ldots, a_n$

for the occurrence of precipitation cells and the cell radius r based on the available point probabilities $p^{(0)}(s_1), \ldots, p^{(0)}(s_n)$ and fit the (conditional) distributions of the scaling variables $C_1, \ldots, C_n$ based on the expectations $\epsilon(s_1), \ldots, \epsilon(s_n)$ and variances $v(s_1), \ldots, v(s_n)$ of point precipitation amounts, where one of the five distributions presented in Section 5.4.2 is used. In order to generate realizations of the random field $\{\Gamma(t), t \in W\}$ of precipitation amounts (conditioned on $\{E = e\}$), we first simulate the random germ-grain model M of precipitation cells (conditioned on $\{E = e\}$) according to Algorithm 3.3.5 and then generate realizations of the random scaling variables $C_1, \ldots, C_n$ (also conditioned on $\{E = e\}$). Simulation algorithms for all five distributions can be found in Kroese et al. (2011), Section 4.2.1 (for the beta prime distribution), Section 4.2.6 (for the gamma distribution), Section 3.1.2.3 (for the inverse gamma distribution), Section 4.2.17 (for the inverse normal distribution), and Section 4.2.10 (for the log-normal distribution). We suggest to use the following Monte Carlo estimator of the (deterministic) area probability $\pi^{(u)}(B) = \mathbb{E}\,(\Pi^{(u)}(B) \mid E = e)$. For an arbitrary integer $j \in \mathbb{N}$ let $\{\Gamma^{(1)}(t), t \in W\}, \ldots, \{\Gamma^{(j)}(t), t \in W\}$ be a sequence of (conditionally) independent and identically distributed random fields in W with $\{\Gamma^{(i)}(t), t \in W\} \stackrel{\mathrm{d}}{=} \{\Gamma(t), t \in W\}$ for $i = 1, \ldots, j$ (conditioned on $E = e$). Then, an estimator $\hat{\pi}^{(u)}(B)$ of $\pi^{(u)}(B)$ is given by

$$\hat{\pi}^{(u)}(B) = \frac{1}{j}\,\#\{i \in \{1, \ldots, j\} : \max\{\Gamma^{(i)}(t), t \in B\} > u\}. \tag{5.10}$$

Similarly, the expected precipitation amount $\epsilon(t)$ for any $t \in W$ can be estimated by the empirical mean of $\Gamma^{(1)}(t), \ldots, \Gamma^{(j)}(t)$. This seems to be more efficient than using (5.6), where a large number of integrals has to be computed numerically.

We present some sample results that were obtained using the proposed model of precipitation amounts (with shape parameter $p = 1$ and gamma distributed scaling variables) based on the point forecast data described in Section 5.1. In Figure 5.6, typical realizations of the random field $\{\Gamma(t), t \in W\}$ of precipitation amounts are shown for two sample forecast periods. One would expect that the radius r of precipitation cells (representing precipitation of any amount) is bigger than the radius computed according to the approach presented in Chapter 4, where only precipitation amounts of more than 0.1 mm are considered. In order to account for this we consider the possible radii $25, 27.5, \ldots, 65$ (in pixel) for all forecast periods in summer 2012 and the radii $30, 32.5, \ldots, 70$ for winter 2012. For forecast period June 16, 2012, 20-21 UTC, however, the radius r is slightly smaller than the one computed in Chapter 4 (27.5 pixel instead of 30), which can most likely be explained by random estimation errors. In forecast period December 9, 2012, 11-12 UTC a larger radius is computed using the approach presented in this chapter (70 instead of 52.5 pixel), which is also observed for most other forecast periods. Although simulated precipitation patterns look atypical compared to radar data, see Figure 4.3, we observe that precipitation cells with high amounts are mainly generated in regions where indeed high precipitation amounts



June 16, 2012, 20-21 UTC, $r = 27.5$ pixel December 9, 2012, 11-12 UTC, $r = 70$ pixel

Figure 5.6.: Realizations of the random field $\{\Gamma(t), t \in W\}$ of precipitation amounts with shape parameter $p = 1$ and gamma distributed scaling variables for two selected forecast periods.

occurred and vice versa. In Figures 5.7 and 5.8, examples of area probabilities for all Voronoi cells $V(s_1), \ldots, V(s_{503})$ that intersect the subset $W_{in} \subset W$ of locations with a certain distance to the boundaries of Germany (see Section 4.5.2) are illustrated. Area probabilities are estimated based on 5,000 realizations of the random field $\{\Gamma(t), t \in W\}$ of precipitation amounts for thresholds $u \in \{0, 0.2, 0.7, 2\}$, which can be done in a reasonable computation time. We again find a clear relationship between estimated area probabilities and underlying point probabilities, compare to Figures 5.2 and 5.3. Area probabilities are always significantly higher than point probabilities but both types of probabilities take lower and higher values in the same regions. For forecast period June 16, 2012, 20-21 UTC we observe that area probabilities for the occurrence of precipitation of more than 0.2 mm are comparable to those estimated for the occurrence of precipitation of more than 0.1 mm in Section 4.6, see Figure 4.10 (left), which can be explained by the smaller cell radius. Finally, Figure 5.9 shows estimated mean precipitation amounts in W for two sample forecast periods. A comparison with expected precipitation amounts $\epsilon(s_1), \ldots, \epsilon(s_{503})$ in Figure 5.4 reveals a close correspondence. In particular, smooth transitions near the boundaries of the Voronoi cells are provided.

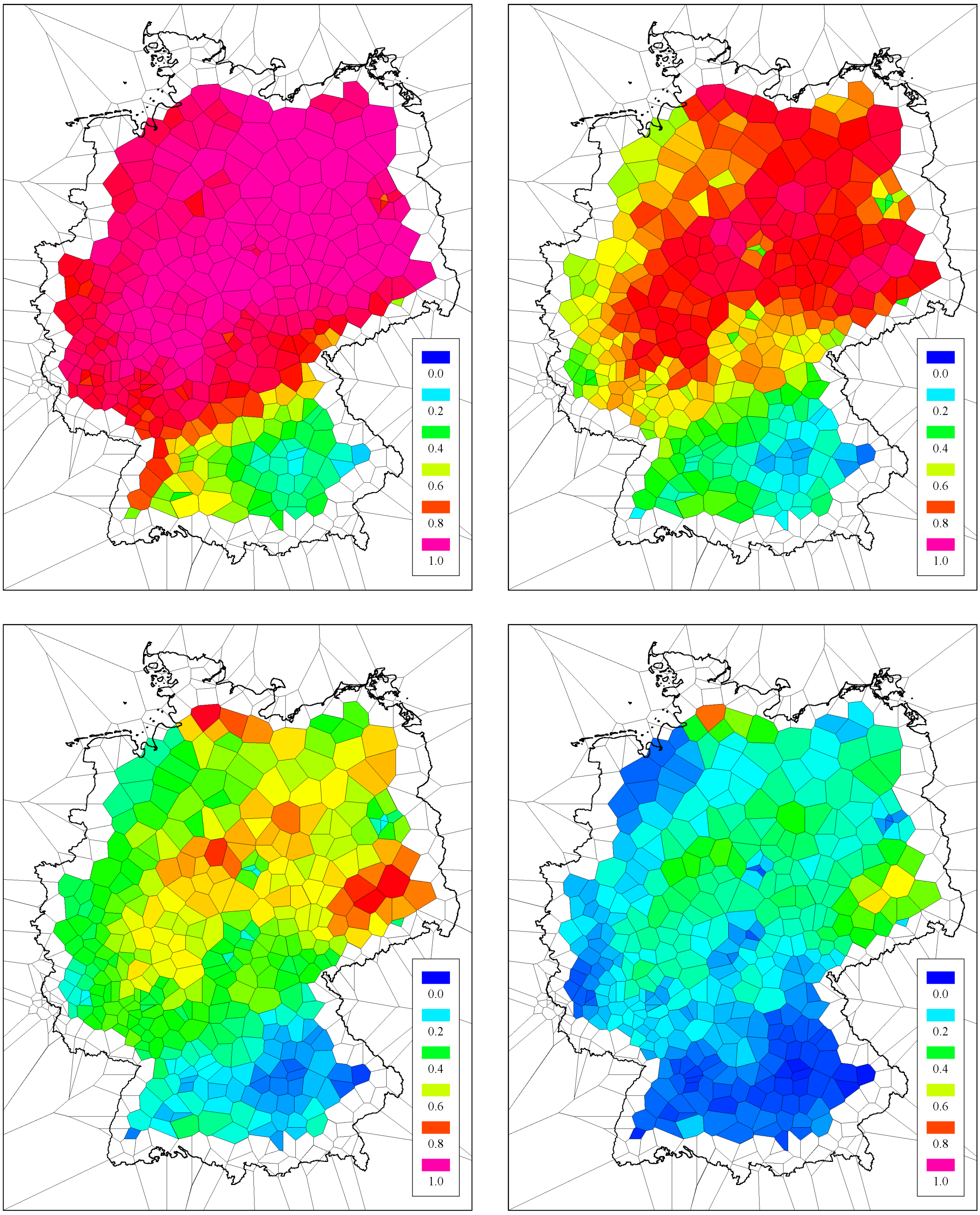


Figure 5.7.: Area probabilities for the occurrence of precipitation of more than 0 mm (top left), 0.2 mm (top right), 0.7 mm (bottom left), and 2 mm (bottom right) in forecast period July 16, 2012, 20-21 UTC. Area probabilities are estimated based on 5,000 realizations of the random field $\{\Gamma(t), t \in W\}$ with cell radius $r = 27.5$ pixels, shape parameter $p = 1$, and gamma distributed scaling variables for all Voronoi cells intersecting W_{in}.

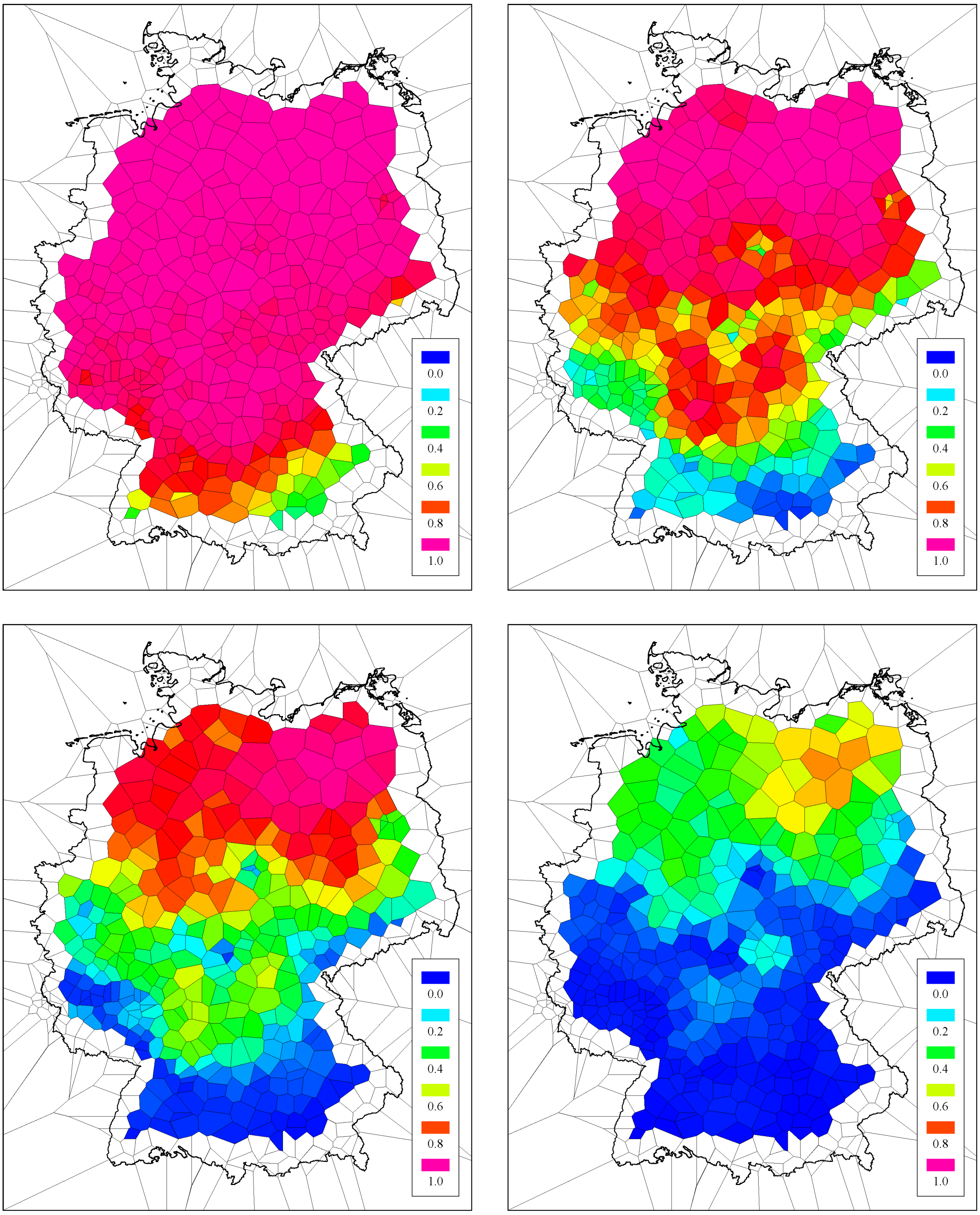
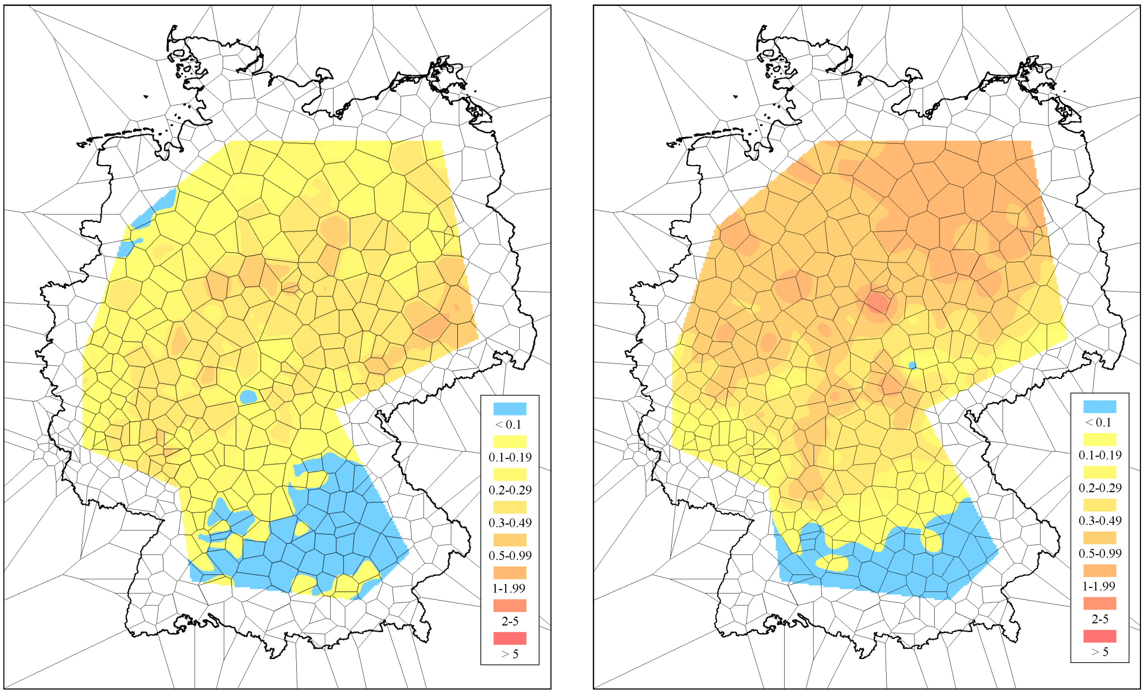


Figure 5.8.: Area probabilities for the occurrence of precipitation of more than 0 mm (top left), 0.2 mm (top right), 0.7 mm (bottom left), and 2 mm (bottom right) in forecast period December 9, 2012, 11-12 UTC. Area probabilities are estimated based on 5,000 realizations of the random field $\{\Gamma(t), t \in W\}$ with cell radius $r = 70$ pixels, shape parameter $p = 1$, and gamma distributed scaling variables for all Voronoi cells intersecting W_{in}.



June 16, 2012, 20-21 UTC, $r = 27.5$ pixel

December 9, 2012, 11-12 UTC, $r = 70$ pixel

Figure 5.9.: Mean precipitation amounts in W_{in} estimated based on 5,000 realizations of the random field $\{\Gamma(t), t \in W\}$ with shape parameter $p = 1$ and gamma distributed scaling variables for two selected forecast periods.

5.6. Forecast verification

To conclude this chapter, a comparison of forecasts with rain gauge adjusted radar measurements is performed using the data described in Section 5.1. However, before applying the proposed model of precipitation amounts to estimate area probabilities for the available $m = 854$ forecast periods, the shape parameter p of the response function κ_p defined in (5.1) and the type of (conditional) distribution of the random scaling variables $C_1, \ldots, C_{503}$ need to be chosen.

5.6.1. Choice of model configuration

We suggest to compare point probabilities obtained by fitting gamma distributions to the available data (see Section 5.3) with point probabilities estimated based on repeated simulation of the proposed model of precipitation (see Section 5.5) to give a recommendation on how to choose the shape parameter p and the type of (conditional) distribution of the local scaling variables $C_1, \ldots, C_{503}$. We consider p to take one of the five values from $\{0.5, 1, 2, 3, 4\}$ and the five types of two-parameter distributions for the scaling variables considered in Section 5.4.2: beta prime, gamma, inverse gamma, inverse normal, and log-normal. It does not seem necessary to consider values $p > 4$ as this does not lead to significant changes in estimated probabilities compared to $p = 4$. For each value of p, each distribution type, and each threshold $u \in \mathbb{T} \setminus \{0\}$ in mm, point probabilities at $s_1, \ldots, s_{503}$ are determined for all available forecast periods using the Monte Carlo estimator (5.10). Then, a comparison with the point probabilities obtained from fitted gamma distributions, see (5.3), is made. For each shape parameter, distribution type, threshold, and weather station, the bias and the mean squared difference (MSD) are computed using the point probabilities of all available forecast periods. Since this results in a huge amount of computed values, the scores are averaged once more over all weather stations. As no consistent estimation of point probabilities can be guaranteed for weather stations outside the boundaries of Germany (due to boundary effects), only stations in W_{in} are taken into account here. The results reveal that for all thresholds and shape parameters, the model performs best regarding MSDs when using gamma distributions for the scaling variables $C_1, \ldots, C_{503}$. However, the effect of changing this type of distribution seems to be minor since only small variations in the scores are observed. Similar results are found when analyzing the effect of changing the shape parameter (when fixing the gamma distribution). The scaled Epanechnikov kernel ($p = 1$) leads to the smallest MSDs in almost all cases but differences are minor for $p = 2, 3, 4$, see Table 5.1, where mean MSDs are listed for thresholds $u \in \{0.1, 0.3, 0.7, 2, 5\}$. Only the surface of the upper half of the 3-dimensional unit ball ($p = 0.5$) produces larger MSDs and significant biases in estimated point probabilities, making $\kappa_{0.5}$ inappropriate for the use as response function. Since the computed scores barely depend on the shape parameter and the distribution type, we also consider a verification of area probabilities using radar data in order to give a final recommendation on how to choose these model configurations, see Section 5.6.2.

5.6.2. Score functions

We perform a forecast verification by comparing estimated area probabilities with precipitation indicators derived from independent rain gauge adjusted radar data. As test areas we again choose the Voronoi cells $V(s_1), \ldots, V(s_{503})$ that correspond to the locations $s_1, \ldots, s_{503}$ of the 503 weather stations since they include areas of different shapes, sizes, and orientations. For each value of p and each of the five distributions suggested in Section 5.4.2, area probabilities for $V(s_1), \ldots, V(s_{503})$ are determined according to the estimator (5.10) for all $m = 854$ forecast periods and all thresholds $u \in \mathbb{T} \setminus \{0\}$ in mm. To each estimated area probability for the occurrence of precipitation of more than u mm somewhere in a Voronoi cell $V(s_i)$ for $i \in \{1, \ldots, 503\}$,

Table 5.1.: Mean MSDs of point probabilities obtained by fitting gamma distributions to the available data and point probabilities estimated based on repeated simulation of the proposed model of precipitation with gamma distributed scaling variables for different shape parameters and thresholds.

threshold in mm	0.1	0.3	0.7	2	5
MSD for $p = 0.5$	$2.6 \cdot 10^{-2}$	$2.5 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	$2.1 \cdot 10^{-3}$	$1.2 \cdot 10^{-4}$
MSD for $p = 1$	$3.3 \cdot 10^{-3}$	$2.1 \cdot 10^{-3}$	$1.16 \cdot 10^{-3}$	$2.7 \cdot 10^{-4}$	$3.2 \cdot 10^{-5}$
MSD for $p = 2$	$3.5 \cdot 10^{-3}$	$2.4 \cdot 10^{-3}$	$1.3 \cdot 10^{-3}$	$2.71 \cdot 10^{-4}$	$3.13 \cdot 10^{-5}$
MSD for $p = 3$	$4.2 \cdot 10^{-3}$	$2.9 \cdot 10^{-3}$	$1.48 \cdot 10^{-3}$	$2.81 \cdot 10^{-4}$	$3.4 \cdot 10^{-5}$
MSD for $p = 4$	$5.1 \cdot 10^{-3}$	$3.4 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	$3 \cdot 10^{-4}$	$4 \cdot 10^{-5}$

we assign the corresponding precipitation indicator, which is 1 if there is precipitation of more than u mm somewhere within $V(s_i)$ with respect to radar data and 0 otherwise. In order to provide a systematic forecast verification for each fixed threshold, we again compute biases, BSSs, and empirical correlation coefficients according to (4.15)-(4.19) based on area probabilities and precipitation indicators of all forecast periods. To further increase the significance of the verification results, all three scores are only determined if the corresponding weather event occurs at least 10 times in the considered Voronoi cell over the entire time period.

At first, we analyze the performance of the three score functions when varying the shape parameter p and the type of (conditional) distribution of the local scaling variables $C_1, \ldots, C_{503}$. The results largely correspond with those obtained in Section 5.6.1. For almost all shape parameters and thresholds, the gamma distribution yields the highest BSSs and correlation coefficients, although varying the type of distribution has a minor effect. A more noticeable impact (particularly on the bias) is observed when changing the value of the shape parameter p. It seems that larger values of p are more appropriate when computing area probabilities for higher thresholds. To obtain a bias that is as close as possible to 0, we recommend to use the scaled Epanechnikov kernel ($p = 1$) for thresholds smaller than 0.2 mm, the scaled biweight kernel ($p = 2$) for thresholds between 0.2 mm and 0.5 mm and the scaled triweight kernel ($p = 3$) for thresholds of at least 0.5 mm. A larger value of p can improve the bias even more for thresholds of more than 1 mm but this will also lead to decreasing BSSs and correlation coefficients and is therefore not recommended. Choosing $p = 0.5$ causes estimated area probabilities for all thresholds to be systematically too low. If one shape parameter needs to be selected to estimate area probabilities for all thresholds based on the same fitted model, then using $p = 2$ seems to be the best compromise.

We analyze the three considered score functions for estimated area probabilities, where the model configurations suggested above (i.e., gamma distributed scaling variables and varying shape parameters for different thresholds) are used. Since obtained area probabilities are expected to depend heavily on the underlying input data, we also provide a comparison of point probabilities at $s_1, \ldots, s_{503}$ (which are computed based on fitted gamma distributions, see Section 5.3) and precipitation indicators derived from radar data. Again, the bias, the BSS, and the empirical correlation coefficient are considered, where scores are only computed for those weather stations at which the corresponding weather event occurs at least 10 times during the selected time period. This implies, however, that no verification of point probabilities for thresholds of 5 mm or higher is possible. Scores for all considered thresholds are visualized using boxplots in Figure 5.10 for point probabilities and in Figure 5.11 for area probabilities. Again, only weather stations located in W_{in} and Voronoi cells intersecting W_{in} are taken into account.

When analyzing mean biases for estimated area probabilities, we find that there is no systematic error for all thresholds up to 5 mm, whereas area probabilities seem to be slightly too low for thresholds of 10 and 15 mm. More variation is observed for single Voronoi cells. Although the bias is close to zero for most areas, we occasionally obtain values reaching up to $-7\,\%$ or $+10\,\%$, see Figure 5.11 (top). Biases are closer to zero for higher thresholds, since in general the corresponding probabilities are smaller. Several reasons causing these biases are conceivable. On the one hand, radar measurements are susceptible to interference that can result in systematic errors for some regions. On the other hand, we already indicated in Section 4.7.1 that biases in estimated area probabilities are induced by biases in the underlying point probabilities (even if these are smaller). Indeed, we observe positive biases of point and area probabilities in northern Germany and small negative biases in southern Germany. Thus, it seems that occasional biases of estimated area probabilities are caused (and slightly amplified) by the underlying point forecasts.

Next, BSSs and empirical correlation coefficients are analyzed. In general, these scores decrease with increasing threshold (for both point and area probabilities), which shows that precipitation events occurring less frequently are more difficult to predict. We find that averaged scores as well as almost all single scores are clearly positive for all thresholds up to $u = 3$ mm, see Figure 5.11 (center and bottom). Furthermore, a direct comparison with the underlying point data, see Figure 5.10 (center and bottom), shows that BSSs and correlation coefficients of estimated area probabilities for thresholds up to 2 mm actually perform slightly better than those of point forecasts. Even the few weather stations with more unreliable point probabilities, indicated by very low (or negative) BSSs and correlation coefficients, do not affect estimated area probabilities very much, which is a particularly nice result. Again, we observe that best results (i.e., highest score functions) for area probabilities are obtained in those regions where the underlying data have the highest scores as well, whereas small insufficiencies in our method seem to be influenced by less reliable data. A meaningful verification of area probabilities estimated for thresholds of 5 mm or higher is difficult since the

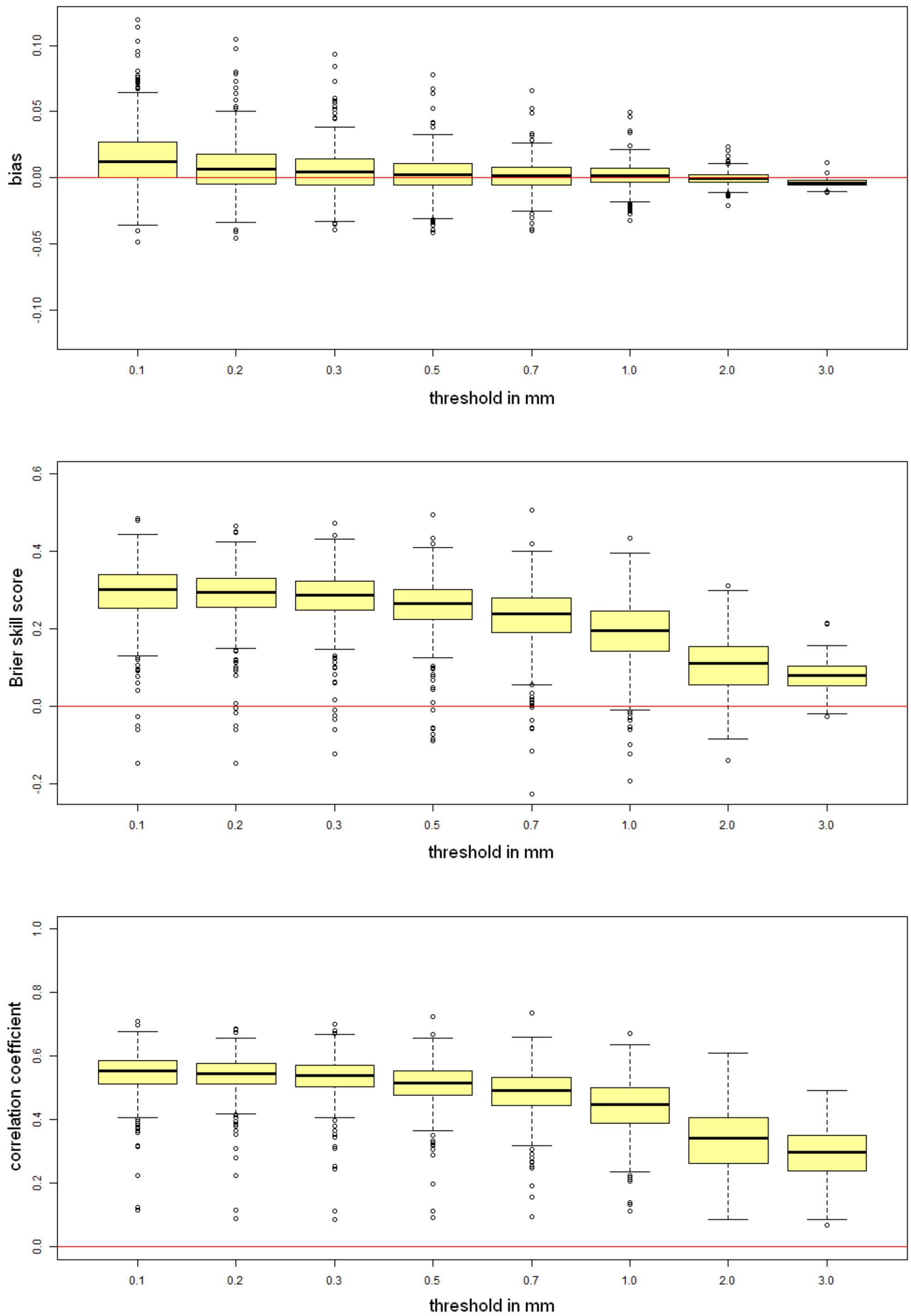


Figure 5.10.: Boxplots showing biases (top), BSSs (center), and empirical correlation coefficients (bottom) of point probabilities that are obtained by fitting gamma distributions to data provided by DWD for all weather stations $s_1, \ldots, s_{503}$ in W_{in}.

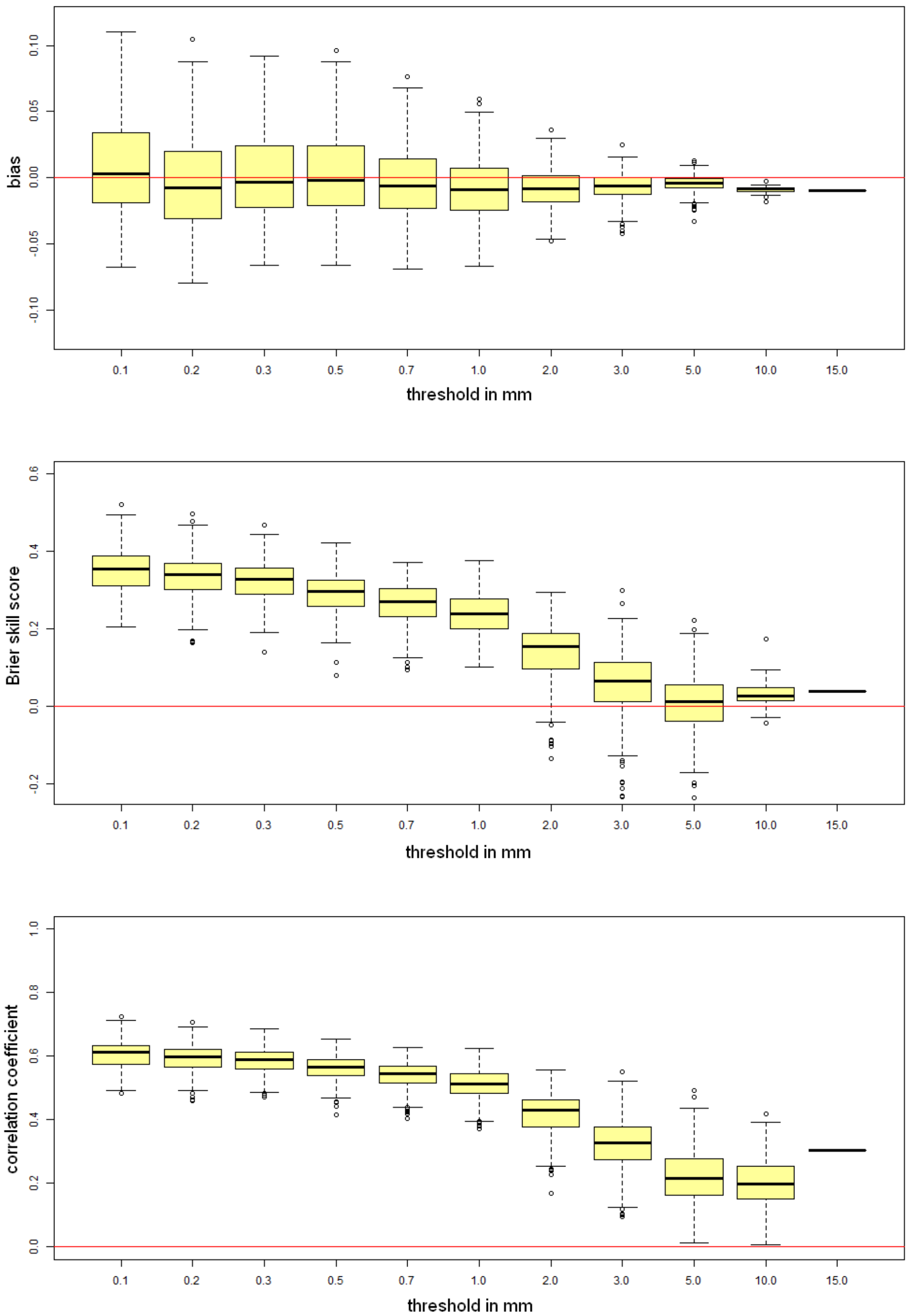


Figure 5.11.: Boxplots showing biases (top), BSSs (center), and empirical correlation coefficients (bottom) of area probabilities that are estimated based on the proposed precipitation model for all Voronoi cells $V(s_1), \ldots, V(s_{503})$ intersecting W_{in}.

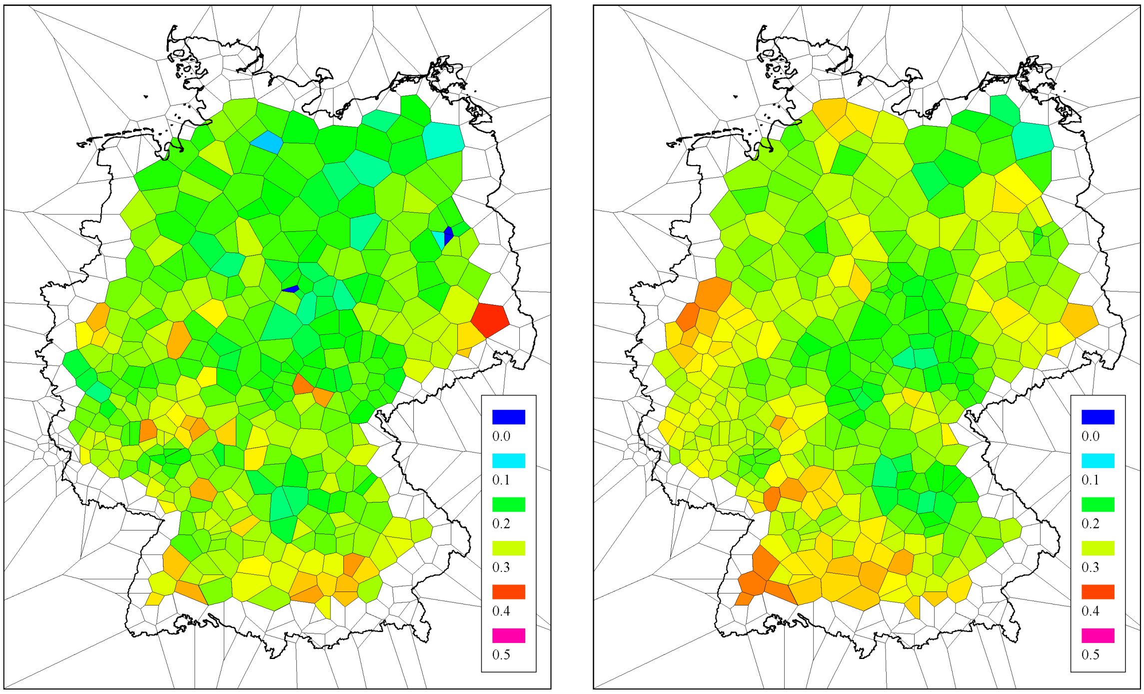


Figure 5.12.: RPSSs of point probabilities that are obtained by fitting gamma distributions to data provided by DWD for weather stations $s_1, \ldots, s_{503}$ (left) and of estimated area probabilities for Voronoi cells $V(s_1), \ldots, V(s_{503})$ (right). RPSSs are only shown for areas intersecting W_{in}.

corresponding (extreme) precipitation events occur rarely in the data. For example, a verification of area probabilities is only possible for 34 Voronoi cells if the threshold is 10 mm and for only 1 Voronoi cell if the threshold is 15 mm. BSSs and correlation coefficients are significantly smaller than for lower thresholds but still positive for most test areas. Although our results indicate that our procedure gives more reliable area probabilities for extreme precipitation events than the climate mean, we also observe that forecast quality is considerably lower than for smaller thresholds. Thus, such area probabilities should be handled with caution.

The score functions computed previously are only able to assess performance of forecasts in dependence of the chosen threshold. To conclude forecast verification, we aim to assess the overall forecast quality of the proposed method. For that purpose, we analyze the ranked probability skill score (RPSS), see, e.g., Daan (1985) or Wilks (2011), Section 8.4.9. This score can be considered as a multiple-category version of the BSS and is constructed as follows. At first, the interval $[0, \infty)$ of all possible precipitation amounts in mm is subdivided into a sequence $U_1 = [0, 0.1], U_2 = (0.1, 0.2], \ldots, U_{11} = (10, 15], U_{12} = (15, \infty)$ of 12 subintervals, whose endpoints correspond to the thresholds considered throughout this chapter. For a fixed forecast period $j \in \{1, \ldots, m\}$ and test area $B \in \mathcal{B}(W)$, let $\pi_j^{(u)}(B)$ denote the probability for the occurrence of precipitation of more than u mm somewhere inside B in period j. Based on this, we determine the sequence $\pi_j(B, U_1), \ldots, \pi_j(B, U_{12})$, where for $i = 1, \ldots, 12$ the probability of a precipitation amount in U_i occurring within B in forecast period j is denoted by $\pi_j(B, U_i)$, i.e., $\pi_j(B, U_1) = 1 - \pi_j^{(0.1)}(B)$, $\pi_j(B, U_2) = \pi_j^{(0.1)}(B) - \pi_j^{(0.2)}(B), \ldots, \pi_j(B, U_{11}) = \pi_j^{(10)}(B) - \pi_j^{(15)}(B)$ and $\pi_j(B, U_{12}) = \pi_j^{(15)}(B)$. Furthermore, we consider the indicator variables $I_j(B, U_1), \ldots, I_j(B, U_{12})$, where $I_j(B, U_i)$ is equal to 1 if the maximal precipitation amount observed within B in period j according to radar data falls into the interval U_i and $I_j(B, U_i)$ is equal to 0 otherwise for $i = 1, \ldots, 12$. Then, the ranked probability score (RPS) rps_π is computed as

$$rps_\pi = \frac{1}{m} \sum_{j=1}^{m} \sum_{k=1}^{12} \left(\sum_{i=1}^{k} \pi_j(B, U_i) - \sum_{i=1}^{k} I_j(B, U_i) \right)^2 .$$

Similar as for the BS, another RPS $rps_{\tilde{\pi}}$ is computed using a reference prediction (where often the climate mean defined in (4.16) is used) and the RPSS $rpss_\pi$ is determined as

$$rpss_\pi = 1 - \frac{rps_\pi}{rps_{\tilde{\pi}}}.$$

RPSSs for the underlying point probabilities can be computed analogously. Values of this score should be positive and as high as possible. Figure 5.12 shows the RPSSs for point and area probabilities, where only cells intersecting W_{in} are colored. All computed values (except for three weather stations) are clearly positive with values between 0.15 and 0.4. In particular, we find that scores for area probabilities have similar values as those for point probabilities, which indicates that our method provides area forecasts with a similar quality as point forecasts included in the (revised) data. The mean RPSS of area probabilities over all Voronoi cells even has a higher value than that for point probabilities over all weather stations (point probabilities: 0.25, area probabilities: 0.28). We conclude that different precipitation amounts occur mainly in those areas and forecast periods, where the corresponding probabilities are high, which shows the success of the developed model-based approach to the estimation of area probabilities for precipitation events.

6. Cluster-based modeling of thunderstorm cells

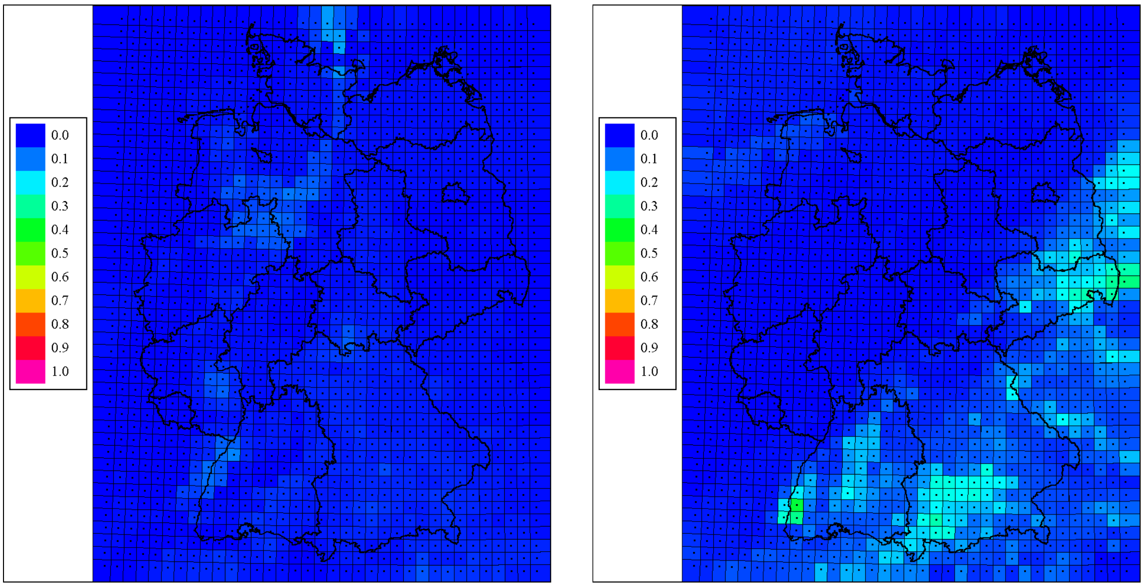
Besides the occurrence of precipitation amounts exceeding certain (warning) thresholds, which is considered in Chapters 4 and 5, other weather events are of great interest in PWP, too. In the present chapter, we address the stochastic modeling and simulation of thunderstorm cells with the goal of estimating reliable point and area probabilities for their occurrence. It is generally recognized that thunderstorm events are much more difficult to predict than, e.g., the occurrence of precipitation, which is why reliable forecasts can usually be given for short forecast ranges only. Furthermore, thunderstorm cells typically have different properties than precipitation cells, making it unclear whether similar stochastic models can be used for both types of weather events. The available data used in this chapter to illustrate the proposed models and simulation algorithms are presented in Section 6.1. They are provided by DWD and include both point forecasts for the occurrence of thunderstorms and historical thunderstorm records. In a first approach, considered in Section 6.2, we suggest the application of the model for precipitation cells introduced in Chapter 4 and demonstrate why it is not suitable for the characterization of thunderstorms. In Section 6.3, a more sophisticated model is developed, which enables a more realistic representation of thunderstorm cells using cluster processes. Several statistical methods for the computation of model characteristics are presented in Section 6.4. The estimation of area probabilities for the occurrence of thunderstorms based on repeated Monte Carlo simulation is discussed in Section 6.5, where we put a particular focus on a conditional simulation algorithm that significantly improves the forecast quality for short ranges by combining point probabilities with thunderstorm records from past periods. Finally, Section 6.6 provides a forecast verification by comparing point and area probabilities that were derived by (conditional) simulation of the proposed model with thunderstorm data. The results of this chapter have been summarized in Kriesche et al. (2017b).

6.1. Description of data

The model-based methods for the estimation of thunderstorm probabilities developed in this chapter combine data from PWP and high-resolution thunderstorm records from past periods. As a basis for choosing a suitable domain, we first consider a rectangular window in central Europe that contains the points $\{(i,j), i,j = 1,\ldots,900\}$ of the RADOLAN lattice introduced in Section 2.1. Recall that point coordinates which refer to the RADOLAN lattice are denoted as RADOLAN coordinates in this thesis and that neighboring lattice points always have a distance of 1 km, which implies that using the Euclidean distance for pairs of RADOLAN coordinates is more precise and produces smaller errors than using the projected window considered in Chapters 4 and 5 (a small, negligible error still occurs due to the curvature of the earth's surface). The window that comprises the RADOLAN lattice is slightly extended to the south, east, and north (to cover the entire territory of Germany and to allow for a more convenient representation) and cut in the west, where large regions of the Netherlands, Belgium, and France are included that are not of interest in our example of application. This finally results in a rectangular window $[150, 925] \times [-25, 950]$ covering the area shown in Figure 6.1, where all locations in this domain are given in RADOLAN coordinates. A conversion of geographical coordinates into RADOLAN coordinates and back can be done using an algorithm provided by DWD. The Euclidean distance between any two locations in $[150, 925] \times [-25, 950]$ (given in RADOLAN coordinates) can be interpreted as the (approximate) distance between the corresponding geographical coordinates in km.

Point forecast data are available for a regularly spaced lattice consisting of 1,575 points inside the domain $[150, 925] \times [-25, 950]$ given in RADOLAN coordinates. The lattice is designed in such a way that neighboring points always have a distance of 20 km. Furthermore, we consider a sequence of 2,205 forecast periods, each with a length of one hour, covering the months May, June, and July 2016 with forecast ranges of one to three hours ahead. This period is of particular interest as in the early summer 2016 an exceptionally high number of severe thunderstorms occurred, including heavy precipitation and hail events. Note that thunderstorms are much more difficult to predict than precipitation, which is why no forecast ranges of more than three hours are considered in our example of application. For each of the selected 2,205 forecast periods, point probabilities for the occurrence of thunderstorms were computed for all 1,575 points of the 20 km $\times$ 20 km lattice using the ModelMIX system of DWD, see Section 2.3.2. In Figure 6.1, point probabilities for two sample forecast periods are illustrated. In general, thunderstorms occur much more infrequently than precipitation, which causes probabilities for thunderstorms to be considerably lower than those for the occurrence of precipitation in most forecast periods. In particular, almost all point probabilities in our available data are smaller than 0.5. In forecast period May 22, 2016, 19-20 UTC, we observe low point probabilities in the entire domain, which suggest the occurrence of only few scattered thunderstorm cells. In the period July 11, 2016, 14-15 UTC, significantly higher point probabilities of up to 0.4 were predicted for some regions of southern and eastern Germany.

As a second source of data historical records of thunderstorm cells from NowCastMIX



May 22, 2016, 19-20 UTC

July 11, 2016, 14-15 UTC

Figure 6.1.: Available point forecast data for two selected forecast periods: the 1,575 points of the considered 20 km $\times$ 20 km lattice and the corresponding Voronoi tessellation, where each Voronoi cell is colored according to the probability for the occurrence of a thunderstorm at the corresponding lattice point.

are available. For each selected forecast period, centers of occurring thunderstorm cells are identified using three different methods, see Section 2.1, and thunderstorm cells are modeled as discs with a globally fixed radius of 10 km. Thunderstorm records from NowCastMIX are given in RADOLAN coordinates, which further motivates choosing the domain as introduced above. Besides the locations of thunderstorm cell centers, NowCastMIX contains a variety of other information. Of particular interest in this chapter are the current wind speed and direction at the location of the cell center as well as a hail flag (an indicator of thunderstorm strength taking the values 0, 1 or 2), which are derived from radar reflectivities using several radar processing methods. Based on these characteristics, a so-called warning cone is computed for each thunderstorm cell using a propagation angle of 7.5°, which reflects the possible movement of the cell in the subsequent one-hour period. Figure 6.2 provides examples of NowCastMIX data for two sample forecast periods. While records for forecast period July 11, 2016, 14-15 UTC correspond well to the point probabilities illustrated in Figure 6.1 (right), we find that in period May 22, 2016, 19-20 UTC, a considerably larger number of thunderstorm cells occurred than was expected due to the relatively low point probabilities, particularly in southern Germany.

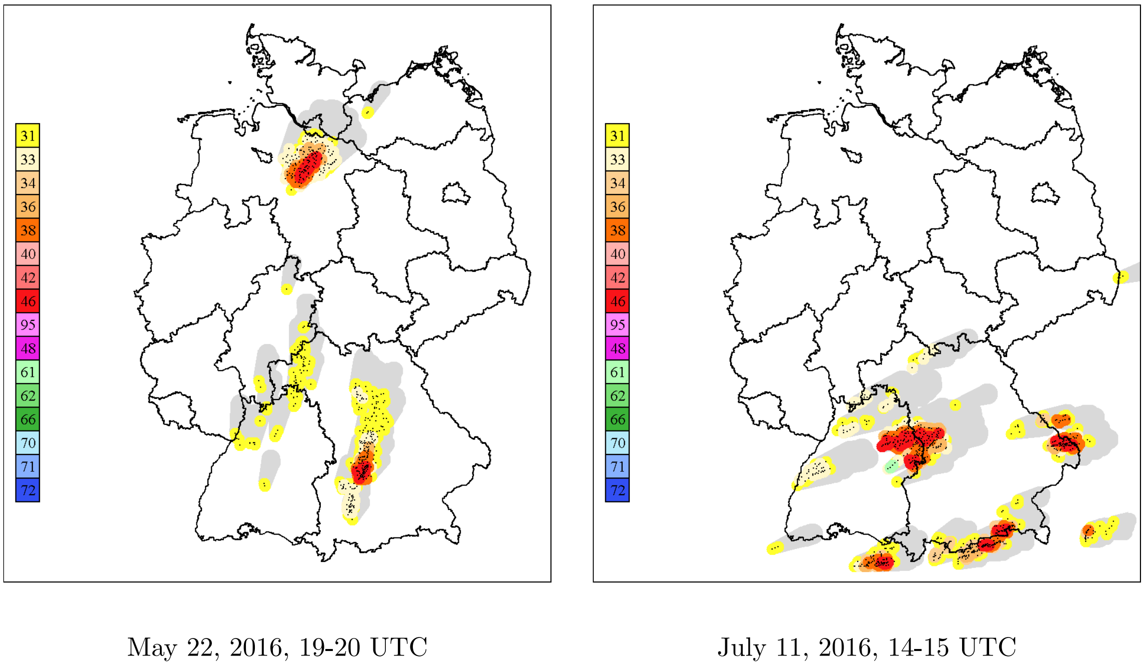
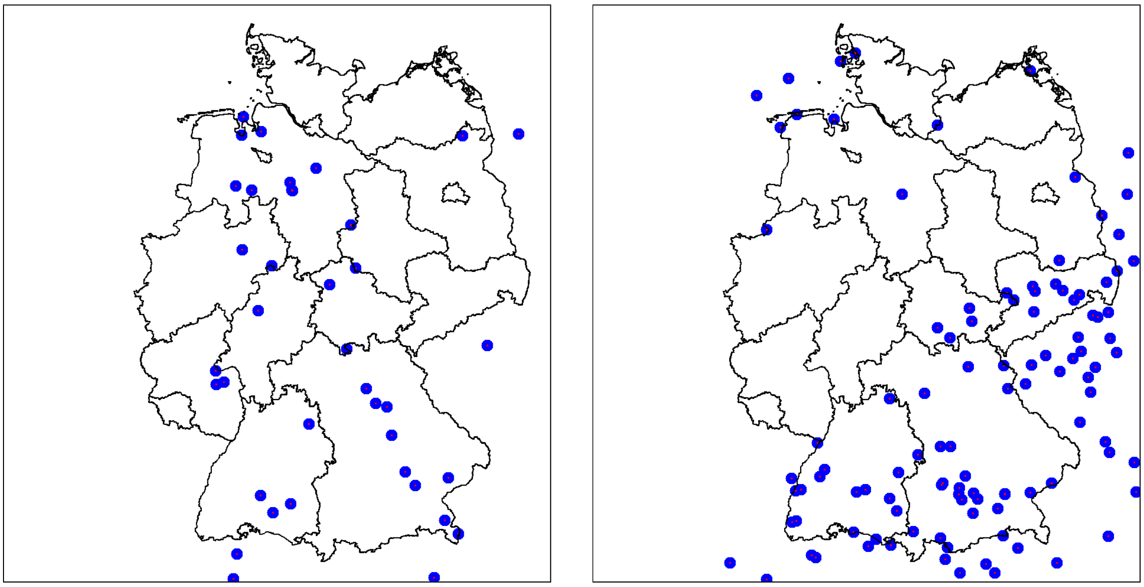


Figure 6.2.: Data from NowCastMIX for two selected forecast periods: recorded thunderstorm cells together with warning cones that reflect the possible movement of cells in the subsequent one-hour period. Thunderstorm cells are colored according to an internal classification of DWD.

6.2. Application of the model for precipitation cells

As a first approach to the modeling of thunderstorm cells we apply the Cox germ-grain model, which was introduced in Chapter 4 for the representation of precipitation cells. To provide consistence of modeled thunderstorm cells with NowCastMIX data, the radius of circular grains is set to a fixed value of 10 km, which means that the time-consuming algorithm presented in Sections 4.5.2 and 4.5.3 does not need to be used. It remains to determine for each forecast period the local intensities for the occurrence of thunderstorms that correspond to the 1,575 Voronoi cells of the 20 km $\times$ 20 km lattice shown in Figure 6.1 according to (4.7), where the lattice points and point probabilities described in Section 6.1 are used. Realizations of the fitted germ-grain model for two sample forecast periods are illustrated in Figure 6.3. Although thunderstorm cells are mainly generated in those regions where positive point probabilities are forecasted and where indeed thunderstorms occurred according to NowCastMIX data, we observe that simulated and recorded thunderstorm patterns have a completely different structure. While thunderstorm cells that are generated according to the proposed Cox germ-grain model appear to be widely scattered, we find that real thunderstorm cells recorded in NowCastMIX seem to occur in clusters. In order to generally assess the applicability



May 22, 2016, 19-20 UTC July 11, 2016, 14-15 UTC

Figure 6.3.: Realizations of the Cox germ-grain model introduced in Chapter 4 for the representation of precipitation cells applied to the modeling of thunderstorm cells with a cell radius of 10 km for two selected forecast periods.

of the Cox germ-grain model for the representation of thunderstorm cells, we analyze the performance of computed area probabilities. This time, the Voronoi cells of the 20 km $\times$ 20 km lattice shown in Figure 6.1 do not appear to be suitable test areas since most of them have the same size and shape. To obtain test areas with more varying shapes, sizes, and orientations, we generate a realization of a stationary Poisson point process in the considered domain $[150, 925] \times [-25, 950]$ using Algorithm 3.3.1, whose intensity is chosen such that the expected number of points is equal to 1,000. We obtain a realization with 999 points and use the cells of the corresponding Voronoi tessellation, denoted as $B_1 \ldots, B_{999}$ in the following, as test areas. Area probabilities for $B_1, \ldots, B_{999}$ can be computed directly according to (4.13) or estimated based on repeated Monte Carlo simulation, see (4.14). For each test area $B \in \{B_1, \ldots, B_{999}\}$, we consider the sequences $\pi_1(B), \ldots, \pi_m(B)$ of area probabilities and $I_1(B), \ldots, I_m(B)$ of thunderstorm indicators, which are 1 if there is a thunderstorm within B with respect to NowCastMIX data and 0 otherwise, for the $m = 2,205$ available forecast periods. In order to systematically compare area probabilities and thunderstorm indicators, we analyze three score functions: the bias, the logarithmic skill score (LSS), and the empirical correlation coefficient. The LSS is preferred over the BSS as this score is recognized to be more suitable for rare weather events such as the occurrence of thunderstorms, see, e.g., Benedetti (2010) or Wilks (2011), Section 8.4.6. For each forecast period $j \in \{1, \dots, m\}$, the ignorance i_j of $\pi_j(B)$ is set to $-\log(\pi_j(B))$ if $I_j(B) = 1$ and to $-\log(1 - \pi_j(B))$ if $I_j(B) = 0$. Then, the logarithmic score (LS) ls_π is defined as the mean of $i_1, \dots, i_m$, i.e.,

$$ls_\pi = \frac{1}{m} \sum_{j=1}^{m} i_j.$$

Since the LS is difficult to compare for different forecast periods, it is related to the LS $ls_{\tilde{\pi}}$ of a reference prediction, where usually the climate mean of $I_1(B), \dots, I_m(B)$ is used, see (4.16). Then, the LSS lss_π is given by

$$lss_\pi = 1 - \frac{ls_\pi}{ls_{\tilde{\pi}}}.$$

Of course, LSs of analyzed area probabilities should not be bigger than those of the (naive) reference prediction, which is why the LSS is requested to be clearly positive. Biases and empirical correlation coefficients can be computed according to (4.15) and (4.19) based on $\pi_1(B), \dots, \pi_m(B)$ and $I_1(B), \dots, I_m(B)$. The three considered score functions are computed and illustrated for all test areas that are not too close to the boundaries of the domain, see Figures 6.4-6.6 (right), where each test area is colored according to the value of the corresponding score function. Since the quality of computed area probabilities is expected to strongly depend on the precision of the underlying point forecast data, the same score functions are also computed for available point probabilities, see Figures 6.4-6.6 (left), where each Voronoi cell is colored according to the value of the score function at the corresponding point of the 20 km $\times$ 20 km lattice. At some locations in the northwest no thunderstorms are recorded for the entire period, which is why the corresponding Voronoi cells are left white.

We find that point probabilities for the occurrence of thunderstorms provided by DWD do not seem to contain any systematic bias (the mean bias is at $-0.2\,\%$, single values range between $-2\,\%$ and $1\,\%$) but only moderate LSSs (mean value of 0.24, most single values between 0 and 0.35) and correlation coefficients (mean value of 0.27, most single values ranging from 0 to 0.4) are obtained. For area probabilities results are different. We get reasonably high correlation coefficients (significantly higher than for point probabilities with values between 0.1 and 0.6 for most test areas), which shows that the proposed model indeed produces higher probabilities in periods and areas where thunderstorms occur than in those where thunderstorms do not occur. However, the biases show that area probabilities are systematically too high. The mean bias is at $3\,\%$ and single values reach up to $7\,\%$, which is large given that the relative frequency of a thunderstorm (i.e., the mean thunderstorm indicator) over all test areas and forecast periods is at $4\,\%$ only. LSSs show slightly smaller values than for point probabilities (mean value of 0.2), with even having negative values for a few test areas. The results of

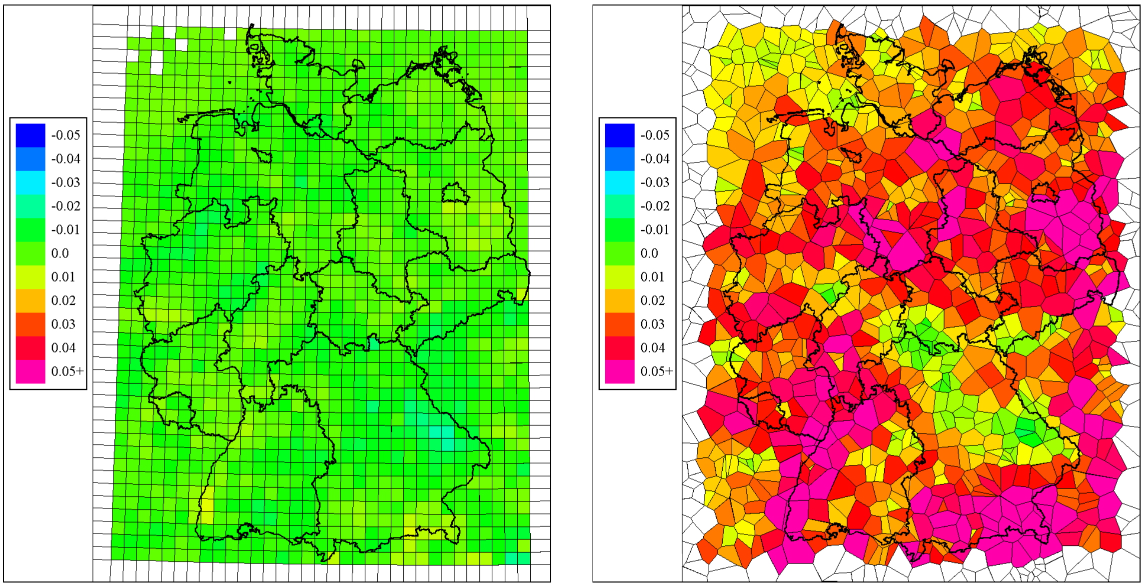


Figure 6.4.: Biases of point probabilities from available data for points of the 20 km $\times$ 20 km lattice (left) and of area probabilities computed according to the Cox germ-grain model introduced in Chapter 4 applied to thunderstorms for test areas $B_1, \ldots, B_{999}$ (right). Biases are not shown for points and areas at the boundaries of the considered domain.

forecast verification (in particular the significant bias) and the comparison of historical thunderstorm records with realizations of the analyzed model demonstrate that the Cox germ-grain model proposed in Chapter 4 for the representation of precipitation cells is not suitable for the modeling of thunderstorms.

6.3. Modeling of thunderstorm cells based on cluster processes

As indicated in Section 6.2, the Cox germ-grain model should not be used for the representation of thunderstorm cells. A probable reason for the model's failure to provide reliable area probabilities is that it generates thunderstorm cells independently of each other in applications, while recorded thunderstorms seem to occur in clusters. Therefore, a different approach for the spatial modeling of thunderstorm cells is proposed in this section. We consider a similar mathematical framework as in Chapters 4 and 5. In the following, let T be a fixed one-hour forecast period that can be interpreted as some subinterval of the real line with a length of 60 minutes. By $T - d$ for $d \geq 0$ we denote the one-hour time period that starts and ends d minutes earlier than T. Furthermore, we consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ introduced

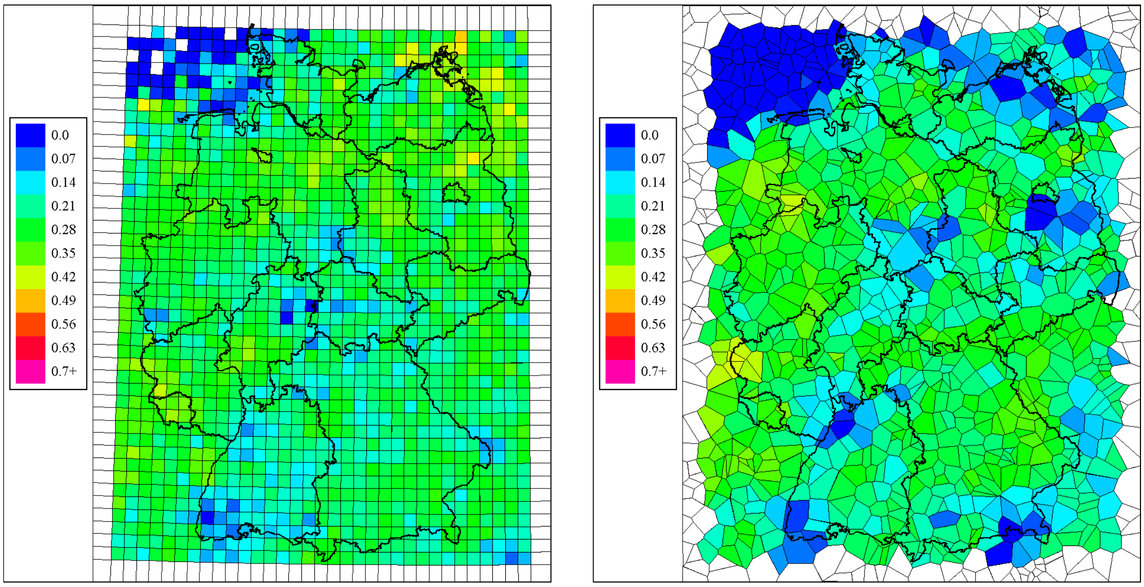


Figure 6.5.: LSSs of point probabilities from available data for points of the 20 km $\times$ 20 km lattice (left) and of area probabilities computed according to the Cox germ-grain model introduced in Chapter 4 applied to thunderstorms for test areas $B_1, \ldots, B_{999}$ (right). LSSs are not shown for points and areas at the boundaries of the considered domain.

in Section 4.2, where the sample space Ω contains all possible weather scenarios in forecast period T and the corresponding forecasts provided by DWD, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and $\mathbb{P}$ is some probability measure on $(\Omega, \mathcal{F})$. The random error occurring in probabilistic forecasts provided by DWD is again modeled by a random element $E : \Omega \to \mathbb{S}$, where $\mathbb{S}$ is the (abstract) measurable space of all such possible errors, and $\sigma(E) \subset \mathcal{F}$ denotes the sub-σ-algebra of events generated by E. We introduce the random field $\{P(t), t \in W\}$ with compact and convex domain $W \subset \mathbb{R}^2$, where the $\sigma(E)$-measurable random variable $P(t) : \Omega \to [0, 1]$ denotes the random probability for the occurrence of a thunderstorm at location $t \in W$ in forecast period T. Similar as in Chapters 4 and 5, we assume that point forecasts are provided by DWD for a sequence of locations $s_1, \ldots, s_n \in W$ with $n \in \mathbb{N}$ and that available data include point probabilities $p(s_1) = \mathbb{E}\,(P(s_1) \,|\, E = e), \ldots, p(s_n) = \mathbb{E}\,(P(s_n) \,|\, E = e)$ for the occurrence of thunderstorms at $s_1, \ldots, s_n$, which are computed based on a particular realization of the weather forecast models with error $e \in \mathbb{S}$. In the example of application considered in this chapter to illustrate the obtained results, we have that $W = [150, 925] \times [-25, 950]$ is a rectangle of projected RADOLAN coordinates covering the territory of Germany and $s_1, \ldots, s_{1575}$ identifies the $n = 1,575$ points of the 20 km $\times$ 20 km lattice illustrated in Figure 6.1. Furthermore, $p(s_1), \ldots, p(s_{1575})$ denote the point probabilities for the occurrence of thunderstorms at $s_1, \ldots, s_{1575}$ described in

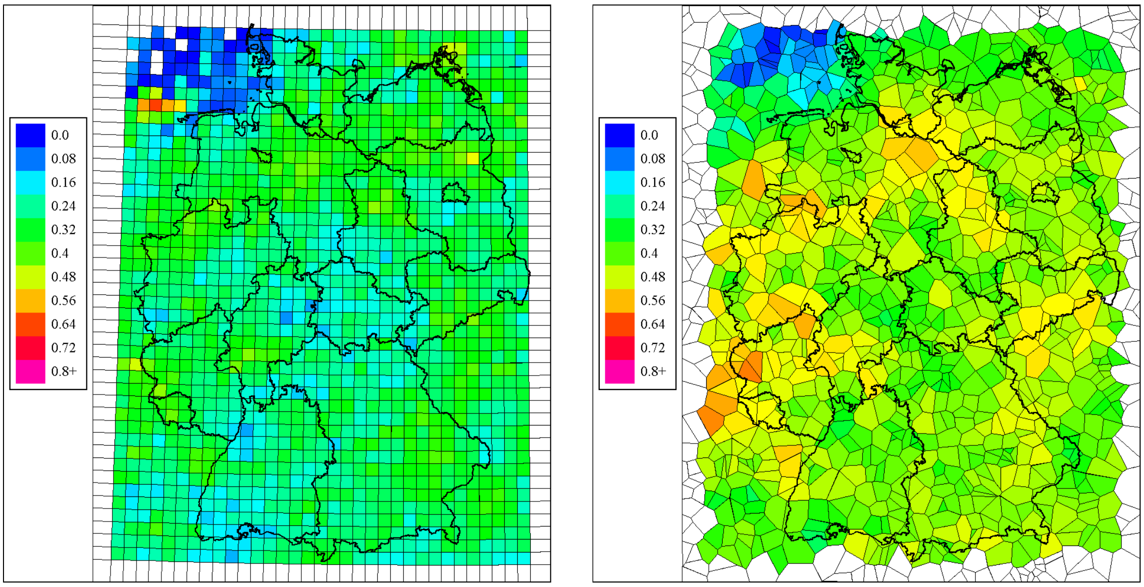


Figure 6.6.: Empirical correlation coefficients of point probabilities from available data for points of the 20 km × 20 km lattice (left) and of area probabilities computed according to the Cox germ-grain model introduced in Chapter 4 applied to thunderstorms for test areas $B_1, \dots, B_{999}$ (right) with corresponding thunderstorm indicators from NowCastMIX data. Correlation coefficients are not shown for points and areas at the boundaries of the considered domain.

Section 6.1 and e is the particular error that occurs when computing these data.

We follow the same fundamental assumption as made in Section 4.3 for the modeling of point probabilities in the context of precipitation. We consider a thunderstorm to occur at some location $t \in W$ if and only if this location is covered by at least one thunderstorm cell. Accordingly, the random field $\{P(t), t \in W\}$ of point probabilities can be represented by

$$P(t) = \mathbb{P}(t \in M \mid E), \quad t \in W, \tag{6.1}$$

where $M : \Omega \to \mathsf{C}$ is a two-dimensional random closed set (see Section 3.3.6) modeling the union set of thunderstorm cells occurring in forecast period T. However, as already shown, the Cox germ-grain model introduced in (4.3) is not a suitable choice for M, which is why a more sophisticated model based on a cluster process is proposed. We start with the modeling of cluster centers. For that purpose, we consider a two-dimensional Cox point process $\{Y_i, i = 1, \dots, Z_Y\}$ in W defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}(Z_Y < \infty) = 1$. We suppose that $\{Y_i, i = 1, \dots, Z_Y\}$ has a random intensity function $\{\Lambda^{(0)}(t), t \in W\}$, where the $\sigma(E)$-measurable random variable $\Lambda^{(0)}(t) : \Omega \to [0, \infty)$ denotes the random intensity for the occurrence of thunderstorm clusters at location $t \in W$. In order to account for the circumstance that point forecasts are only available at the locations $s_1, \ldots, s_n$ but to still enable a spatially inhomogeneous distribution of thunderstorm cells, we suppose that $\{\Lambda^{(0)}(t), t \in W\}$ can be represented by

$$\Lambda^{(0)}(t) = \sum_{j=1}^{n} A_j^{(0)} \mathbb{1}_{V(s_j)}(t), \quad t \in \bigcup_{i=1}^{n} V(s_i), \tag{6.2}$$

where $\{V(s_1), \ldots, V(s_n)\}$ denotes the Voronoi tessellation of $s_1, \ldots, s_n$ defined in (4.2). The $\sigma(E)$-measurable random variables $A_1^{(0)}, \ldots, A_n^{(0)} : \Omega \to [0, \infty)$ can be interpreted as local random intensities for the occurrence of thunderstorm clusters in neighborhoods of $s_1, \ldots, s_n$. If $t \in W$ is located on the boundaries of one or more Voronoi cells, i.e., if $t \notin \bigcup_{i=1}^{n} V(s_i)$, then $\Lambda^{(0)}(t)$ is set to the minimum intensity of all adjacent Voronoi cells.

For the modeling of the clusters themselves some further simplification is necessary. It can be observed that the shapes of recorded thunderstorm clusters vary significantly across space and time and can therefore hardly be determined, see, e.g., the sample data illustrated in Figure 6.2. We suggest to model thunderstorm clusters as circular discs around the points of the Cox process $\{Y_i, i = 1, \ldots, Z_Y\}$ as described in the following. Consider a sequence $X^{(1)} = \{X_i^{(1)}, i = 1, \ldots, Z_{X^{(1)}}\}, X^{(2)} = \{X_i^{(2)}, i = 1, \ldots, Z_{X^{(2)}}\}, \ldots$ of identically distributed Cox point processes in W with $\mathbb{P}(Z_{X^{(i)}} < \infty) = 1$ for all $i \in \mathbb{N}$. We suppose that $X^{(1)}, X^{(2)}, \ldots$ are conditionally independent of each other and of the Cox process $\{Y_i, i = 1, \ldots, Z_Y\}$ of cluster centers given $\{E = e\}$ for any $e \in \mathbb{S}$. To provide circular thunderstorm clusters, we assume that the (identically distributed) Cox processes $X^{(1)}, X^{(2)}, \ldots$ have a random intensity function $\{\Lambda^{(1)}(t), t \in W\}$ defined by

$$\Lambda^{(1)}(t) = A^{(1)} \mathbb{1}_{b(o, R^{(1)})}(t), \quad t \in W, \tag{6.3}$$

where the $\sigma(E)$-measurable random variable $A^{(1)} : \Omega \to [0, \infty)$ can be interpreted as random cluster intensity and the $\sigma(E)$-measurable random variable $R^{(1)} : \Omega \to (0, \infty)$ describes the cluster radius. In order to give a proper representation of the process of all thunderstorm centers, we also consider the random counting measures $\{N^{(1)}(B), B \in \mathcal{B}(W)\}, \{N^{(2)}(B), B \in \mathcal{B}(W)\}, \ldots$ that correspond to the Cox processes $X^{(1)}, X^{(2)}, \ldots$, i.e., $N^{(j)}(B) : \Omega \to \mathbb{N}_0$ with $N^{(j)}(B) = \#\{i : X_i^{(j)} \in B\}$ for all $B \in \mathcal{B}(W)$ and $j \in \mathbb{N}$. Based on this, we introduce the random counting measure $\{N(B), B \in \mathcal{B}(W)\}$ defined by

$$N(B) = \sum_{i=1}^{Z_Y} N^{(i)}((B - Y_i) \cap W), \quad B \in \mathcal{B}(W).$$

Then, according to Lemma 3.3.1, there is a uniquely defined point process $\{X_i, i = 1, \ldots, Z_X\}$ in W with $\mathbb{P}(Z_X < \infty) = 1$ and $N(B) = \#\{i : X_i \in B\}$ for all $B \in \mathcal{B}(W)$, which is used as a model for the centers of thunderstorm cells in forecast period T. This kind of process is referred to as a (doubly-stochastic) cluster process, see Section 3.3.5. Due to the Cox process $\{Y_i, i = 1, \ldots, Z_Y\}$ of cluster centers being non-stationary, the cluster process $\{X_i, i = 1, \dots, Z_X\}$ of thunderstorm cell centers has this property, too.

Finally, we again suggest to model thunderstorm cells as discs $b(X_1, R), \dots, b(X_{Z_X}, R)$ with a random $\sigma(E)$-measurable radius $R : \Omega \to (0, \infty)$ around the locations of the cluster process $\{X_i, i = 1, \dots, Z_X\}$ and represent the union set M of thunderstorm cells as a germ-grain model by

$$M = \bigcup_{i=1}^{Z_X} b(X_i, R). \tag{6.4}$$

Both the cluster process $\{X_i, i = 1, \dots, Z_X\}$ of thunderstorm cell centers and the germ-grain model M of thunderstorm cells cannot be assumed to be $\sigma(E)$-measurable since given $\{E = e\}$ for any realization e of E (i.e., given a specific forecast provided by the weather forecast models of DWD) the actual weather in the future forecast period T is still considered to be random.

6.4. Computation of model characteristics

Since the approach to the spatial stochastic modeling of thunderstorm cells proposed in Section 6.3 is based on a more complex point process model than the one considered in Chapter 4 for the representation of precipitation cells, there are also more model characteristics that need to be determined based on data provided by DWD. In applications, fitting and simulation of the developed models is done based on a particular forecast provided by the weather forecast models of DWD, which is why a fixed realization e of E is considered in the rest of this chapter. Given $\{E = e\}$, the conditional distribution of the cluster-based germ-grain model M specified in (6.4) is completely characterized by the corresponding realizations $a_1^{(0)} = \mathbb{E}\,(A_1^{(0)} \,|\, E = e), \dots, a_n^{(0)} = \mathbb{E}\,(A_n^{(0)} \,|\, E = e)$, $a^{(1)} = \mathbb{E}\,(A^{(1)} \,|\, E = e)$, $r^{(1)} = \mathbb{E}\,(R^{(1)} \,|\, E = e)$, and $r = \mathbb{E}\,(R \,|\, E = e)$ of the random local intensities $A_1^{(0)}, \dots, A_n^{(0)}$ for the occurrence of thunderstorm clusters, the random cluster intensity $A^{(1)}$, the random cluster radius $R^{(1)}$, and the random radius R of thunderstorm cells. Furthermore, we assume that point probabilities $p(s_1) = \mathbb{E}\,(P(s_1) \,|\, E = e), \dots, p(s_n) = \mathbb{E}\,(P(s_n) \,|\, E = e)$ for the occurrence of thunderstorms at sites $s_1, \dots, s_n$ and thunderstorm records from NowCastMIX for past periods are available as data input. In order to provide comparability of simulated thunderstorm cells and recorded cells from NowCastMIX, we set $r = 10$ km in applications as performed in Section 6.6.

6.4.1. Local intensities for the occurrence of thunderstorm clusters

The computation of the local intensities $a_1^{(0)}, \ldots, a_n^{(0)}$ for the occurrence of thunderstorm clusters can be done analogously to the computation of the local intensities for the occurrence of precipitation cells considered in Section 4.5.1. We suggest to derive a representation formula for point probabilities according to the model for thunderstorm cells proposed in Section 6.3 and choose $a_1^{(0)}, \ldots, a_n^{(0)}$ in such a way that point probabilities computed according to the model are as close as possible to those included in the data provided by DWD. In the following, we assume that the cluster intensity $a^{(1)}$, the cluster radius $r^{(1)}$, and the cell radius r are known. An algorithm for the computation of $a^{(1)}$ and $r^{(1)}$ based on NowCastMIX data is developed in Section 6.4.2 and the cell radius r is set to a fixed value of 10 km in our example of application.

Conditioned on $\{E = e\}$, the random point process $\{Y_i, i = 1, \ldots, Z_Y\}$ of cluster centers is a Poisson point process with (deterministic) intensity function $\{\lambda^{(0)}(t), t \in W\}$, where $\lambda^{(0)}(t) = \mathbb{E}\,(\Lambda^{(0)}(t) \mid E = e)$ for $t \in W$. Furthermore, $X^{(1)}, X^{(2)}, \ldots$ is a sequence of independent and identically distributed Poisson point processes with intensity function $\{\lambda^{(1)}(t), t \in W\}$, where $\lambda^{(1)}(t) = \mathbb{E}\,(\Lambda^{(1)}(t) \mid E = e)$ for $t \in W$. Since $X^{(1)}, X^{(2)}, \ldots$ are supposed to be independent of $\{Y_i, i = 1, \ldots, Z_Y\}$ given $\{E = e\}$, the point process $\{X_i, i = 1, \ldots, Z_X\}$ forms a Matérn cluster process (conditioned on $\{E = e\}$) with random intensity function $\{\Lambda(t), t \in W\}$ defined by

$$\Lambda(t) = a^{(1)} \sum_{i=1}^{Z_Y} \mathbb{1}_{b(Y_i, r^{(1)})}(t), \quad t \in W,$$

compare to Example 3.3.1. Due to the properties of $\{X_i, i = 1, \ldots, Z_X\}$ being a Cox process and $\{Y_i, i = 1, \ldots, Z_Y\}$ being a Poisson process conditioned on $\{E = e\}$, a specific representation formula of the point probability $p(t) = \mathbb{E}\,(P(t) \mid E = e)$ for the occurrence of a thunderstorm at location $t \in W$ can be computed. In order to improve readability, we introduce for each $i \in \{1, \ldots, n\}$ the function $\tilde{I}_i : W \to [0, \infty)$, which is defined by

$$\tilde{I}_i(t) = \int_{V(s_i)} 1 - \exp\left\{-a^{(1)} \nu_2\left(W \cap b(t, r) \cap b(z, r^{(1)})\right)\right\} \mathrm{d}z, \quad t \in W.$$

Then, based on the representation of point probabilities as coverage probabilities of the germ-grain model M, see (6.1), we get that

$$\begin{aligned} p(t) &= \mathbb{P}(t \in M \mid E = e) \\ &= 1 - \mathbb{P}(\#\{i : X_i \in b(t, R)\} = 0 \mid E = e) \\ &= 1 - \mathbb{E}\left(\exp\left\{-\int_{W \cap b(t,r)} a^{(1)} \sum_{i=1}^{Z_Y} \mathbb{1}_{b(Y_i, r^{(1)})}(z)\,\mathrm{d}z\right\}\right) \end{aligned}$$

$$= 1 - \mathbb{E}\left(\prod_{i=1}^{Z_Y} \exp\left\{-a^{(1)}\nu_2\left(W \cap b(t,r) \cap b(Y_i, r^{(1)})\right)\right\}\right)$$
$$= 1 - \exp\left\{\int_W \left(\exp\left\{-a^{(1)}\nu_2\left(W \cap b(t,r) \cap b(z, r^{(1)})\right)\right\} - 1\right)\sum_{i=1}^{n} a_i^{(0)} \mathbb{1}_{V(s_i)}(z)\,\mathrm{d}z\right\}$$
$$= 1 - \exp\left\{-\sum_{i=1}^{n} a_i^{(0)} \tilde{I}_i(t)\right\}, \quad t \in W, \tag{6.5}$$

where in the third equality we use the distributional properties of Cox processes, see Definition 3.3.10, and in the fifth equality a representation formula of the probability generating functional of Poisson processes is applied, see (3.24) with $f(z) = \exp\left\{-a^{(1)}\nu_2\left(W \cap b(t,r) \cap b(z, r^{(1)})\right)\right\}$. Accordingly, the cluster intensities $a_1^{(0)}, \ldots, a_n^{(0)}$ should satisfy

$$\log\left(\frac{1}{1 - p(s_j)}\right) = \sum_{i=1}^{n} a_i^{(0)} \tilde{I}_i(s_j), \quad j = 1, \ldots, n,$$

for fixed $a^{(1)} \geq 0$ and $r^{(1)}, r > 0$, which describes a system of n linear equations with the n unknowns $a_1^{(0)}, \ldots, a_n^{(0)} \geq 0$. Due to the constraint of $a_1^{(0)}, \ldots, a_n^{(0)}$ being nonnegative, there is no exact solution of this system of equations in general, which is why we compute $a_1^{(0)}, \ldots, a_n^{(0)}$ in a nonnegative least squares sense according to

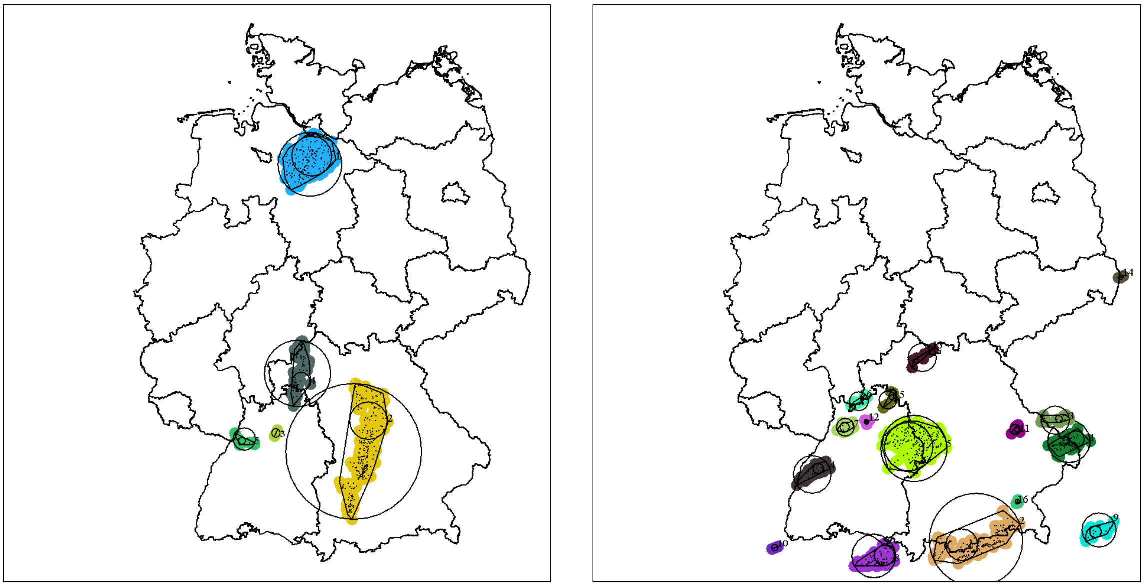
$$(a_1^{(0)}, \ldots, a_n^{(0)}) = \underset{a'_1, \ldots, a'_n \geq 0}{\operatorname{argmin}} \left\{\sum_{j=1}^{n}\left(\log\left(\frac{1}{1 - p(s_j)}\right) - \sum_{i=1}^{n} a'_i \tilde{I}_i(s_j)\right)^2\right\}.$$

6.4.2. Cluster intensity and cluster radius

In order to fully specify the proposed model of thunderstorm cells, it remains to determine the intensity $a^{(1)}$ of cells occurring in each cluster as well as the cluster radius $r^{(1)}$. Unfortunately, it does not seem possible to compute these characteristics as a function of the available point probabilities, too. Thus, we suggest to use thunderstorm records of past periods obtained from NowCastMIX data, see Section 6.1, for this purpose. In the presented fitting approach, clusters of recorded thunderstorm cells are first identified using an established cluster algorithm and are then used to determine $a^{(1)}$ and $r^{(1)}$ based on the sizes of obtained clusters and the total number of recorded thunderstorm cells. Of course, when making a forecast for the one-hour forecast period T, the thunderstorm records for that period are not yet available. Instead, we consider the latest one-hour period prior to T with available NowCastMIX data, which can be represented as $T - d$ with some $d \geq 60$ minutes. The best case ($d = 60$) is given when the forecast is made directly at the beginning of period T and NowCastMIX records of the preceding 60 minutes are already available. While the positions of thunderstorm cells can change quickly over time due to strong wind, we observe that typical sizes of thunderstorm clusters and the number of storms per cluster only change gradually. Thus, we suppose that both $a^{(1)}$ and $r^{(1)}$ (for period T) can be estimated based on NowCastMIX records of period $T - d$ if d is not too large.

At first, a cluster analysis is performed to identify clusters of thunderstorm cell centers in forecast period $T - d$. For this purpose, we suggest to use the Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm introduced in Ester et al. (1996). This algorithm seems to be particularly suitable since it can recognize clusters of arbitrary shapes, it is possible to account for outliers (which are interpreted as noise), and the number of clusters to be found does not need to be known a priori (as required, e.g., in the k-means clustering algorithm). DBSCAN has two parameters: the maximum neighborhood radius ε and the minimum number $minPts$ that is required to form a cluster (clusters with less than $minPts$ points are considered as noise). Comparisons of results for different parameter configurations have shown that when applied to thunderstorm records from NowCastMIX, $\varepsilon = 20$ km and $minPts = 3$ seem to be reasonable choices. Example results for the application of the DBSCAN algorithm to thunderstorm records of two sample forecast periods are illustrated in Figure 6.7, where all thunderstorm cells in one cluster have the same color. A comparison with Figure 6.2 shows that we obtain reasonable partitions of thunderstorm cell centers into clusters and that single isolated cells are indeed identified as noise and thus dropped before computing cluster parameters.

Next, we propose an algorithm to determine $a^{(1)}$ and $r^{(1)}$ based on obtained thunderstorm clusters. Let $c_1, \ldots, c_m$ for some $m \in \mathbb{N}$ denote the clusters of thunderstorm cell centers detected by the DBSCAN algorithm as described above. In order to find typical cluster sizes, we determine for each $i \in \{1, \ldots, m\}$ the radius $r_i^{max} > 0$ of the smallest circle that contains all thunderstorm cell centers of cluster c_i. For detected clusters with an approximately circular shape, e.g., the blue cluster in northern Germany in forecast period May 22, 2016, 19-20 UTC or the light green cluster in southern Germany in period July 11, 2016, 14-15 UTC, see Figure 6.7, a disc with radius r_i^{max} seems suitable. However, for more elongated clusters such as the orange one in southern Germany in forecast period May 22, 2016, 19-20 UTC or the brown one over the Alps in period July 11, 2016, 14-15 UTC, it is unlikely that the cluster can be represented by a disc with radius r_i^{max}. It seems more realistic that such clusters can be modeled by several circular discs with smaller radii that are located directly next to each other. To account for this, we determine for each $i \in \{1, \ldots, m\}$ the convex hull h_i of cluster c_i, which is the smallest convex set that contains all cell centers belonging to c_i, and determine the radius $r_i^{min} > 0$ of the biggest circle that is completely contained in h_i as also illustrated in Figure 6.7. Then, to each cluster c_i for $i \in \{1, \ldots, m\}$ a radius $r_i > 0$ is assigned in dependence of the ratio of r_i^{min} and r_i^{max} according to the following algorithm.



May 22, 2016, 19-20 UTC, 5 clusters

July 11, 2016, 14-15 UTC, 16 clusters

Figure 6.7.: Results of the DBSCAN algorithm with $\varepsilon = 20$ km and $minPts = 3$ applied to thunderstorm records from NowCastMIX data for two selected forecast periods. Additionally, for each cluster the convex hull, the biggest circle contained in the convex hull, and the smallest circle containing all points of the cluster are shown.

1. The minimum cluster radius is supposed to be equal to 10 km. This implies that if $r_i^{max} \leq 10$ km, then $r_i = 10$ km.

2. If 10 km $< r_i^{max} \leq 20$ km, then we always put $r_i = r_i^{max}$.

3. If 20 km $< r_i^{max} \leq 35$ km, then the following applies. If $r_i^{min} < 0.4\, r_i^{max}$, we put $r_i = r_i^{min}$, otherwise $r_i = r_i^{max}$.

4. If 35 km $< r_i^{max} \leq 50$ km, then the following applies. If $r_i^{min} < 0.55\, r_i^{max}$, we put $r_i = r_i^{min}$, otherwise $r_i = r_i^{max}$.

5. If 50 km $< r_i^{max}$, then the following applies. If $r_i^{min} < 0.65\, r_i^{max}$, we put $r_i = r_i^{min}$, otherwise $r_i = r_i^{max}$.

6. The maximum cluster radius is assumed to be equal to 70 km. This implies that if $r_i > 70$ km, then r_i is reduced to 70 km.

Finally, we compute the typical cluster radius $r^{(1)}$ as the mean value of the individual cluster radii $r_1, \ldots, r_m$, i.e.,

$$r^{(1)} = \frac{1}{m} \sum_{i=1}^{m} r_i.$$

To conclude model fitting, the cluster intensity $a^{(1)}$ has to be specified. Let $k \in \mathbb{N}$ denote the total number of thunderstorms contained in all clusters $c_1, \ldots, c_m$ (i.e., thunderstorm centers interpreted as noise by the DBSCAN algorithm are not taken into account). For each $i \in \{1, \ldots, m\}$, we determine the minimal number l_i of discs with radius $r^{(1)}$ that is needed to cover all thunderstorm cell centers contained in cluster c_i. The sum $l = l_1 + \ldots + l_m$ can then be interpreted as the total number of circular clusters with radius $r^{(1)}$ and the ratio k/l denotes the mean number of thunderstorm cells per cluster. Accordingly, the intensity $a^{(1)}$ can be computed as

$$a^{(1)} = \frac{k}{l\,\pi(r^{(1)})^2}.$$

In periods with very weak or no thunderstorm activity, it may happen that no thunderstorms are recorded or that no clusters are detected by the DBSCAN algorithm (which can happen if all thunderstorm cells are considered to be noise) in period $T - d$. In this case, we recommend to put $r^{(1)} = 11$ km and $a^{(1)} = 4/(\pi(r^{(1)})^2) \approx 0.0105$. These are the mean values of $a^{(1)}$ and $r^{(1)}$ from all one-hour periods with thunderstorm activity but without detected thunderstorm clusters in the preceding one-hour period according to the dataset introduced in Section 6.1.

6.5. Monte Carlo simulation of thunderstorm cells

After the proposed model for thunderstorm cells is fully specified, it can be used for the probabilistic prediction of thunderstorm events. For that purpose, we again provide methods for the estimation of area probabilities based on repeated (conditional) Monte Carlo simulation, which can be done in a similar way as for precipitation events, see Sections 4.6 and 5.5.

6.5.1. Model-based computation and estimation of area probabilities

At first, we derive a direct computation formula for area probabilities based on the germ-grain model M of thunderstorm cells proposed in (6.4). We suppose that a thunderstorm occurs somewhere inside a test area $B \in \mathcal{B}(W)$ if and only if B intersects M. Analogous to the derivation of (6.5), we obtain the following representation formula for the (conditional) area probability $\pi(B) \in [0, 1]$ of B (given $\{E = e\}$) using the distributional properties of Cox processes and the formula for the probability generating functional of Poisson processes in (3.24):

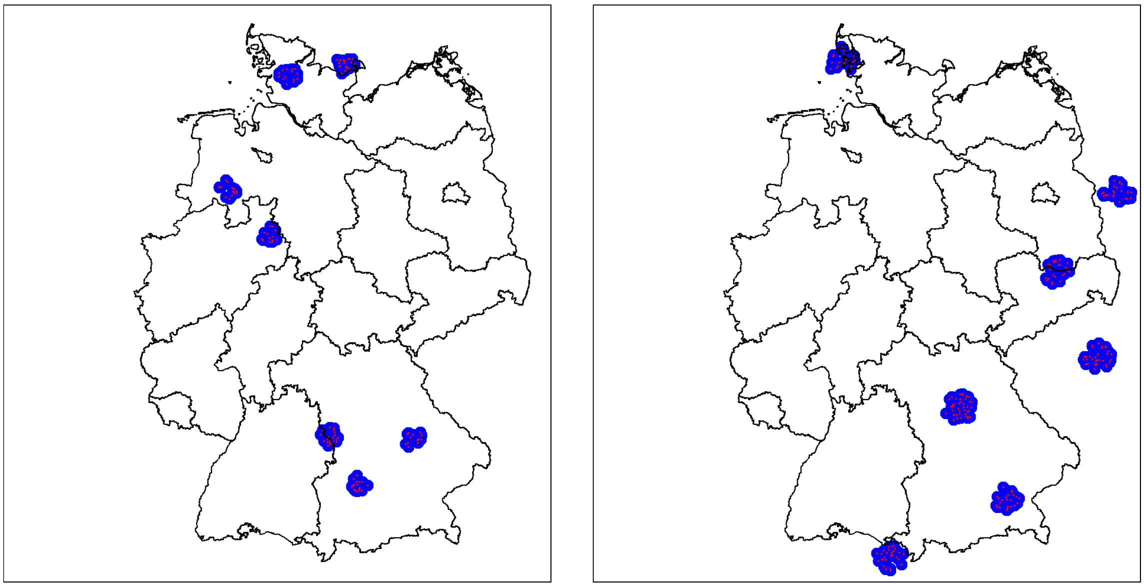
$$
\begin{aligned}
\pi(B) &= \mathbb{P}(B \cap M \neq \emptyset \,|\, E = e)\\
&= 1 - \mathbb{P}(\#\{i : X_i \in B \oplus b(o,R)\} = 0 \,|\, E = e)\\
&= 1 - \mathbb{E}\left(\exp\left\{-\int_{W \cap (B \oplus b(o,r))} a^{(1)} \sum_{i=1}^{Z_Y} \mathbb{1}_{b(Y_i, r^{(1)})}(t)\, \mathrm{d}t\right\}\right)\\
&= 1 - \mathbb{E}\left(\prod_{i=1}^{Z_Y} \exp\left\{-a^{(1)} \nu_2\left(W \cap (B \oplus b(o,r)) \cap b(Y_i, r^{(1)})\right)\right\}\right)\\
&= 1 - \exp\left\{\int_W \left(\exp\left\{-a^{(1)} \nu_2\left(W \cap (B \oplus b(o,r)) \cap b(t, r^{(1)})\right)\right\} - 1\right) \sum_{i=1}^{n} a_i^{(0)} \mathbb{1}_{V(s_i)}(t)\, \mathrm{d}t\right\}\\
&= 1 - \exp\left\{-\sum_{i=1}^{n} a_i^{(0)} \int_{V(s_i)} 1 - \exp\left\{-a^{(1)} \nu_2\left(W \cap (B \oplus b(o,r)) \cap b(t, r^{(1)})\right)\right\}\, \mathrm{d}t\right\}.
\end{aligned}
\tag{6.6}
$$

Alternatively, area probabilities can again be estimated based on repeated Monte Carlo simulation. In most applications, this turns out to be more efficient than using the direct computation formula since a large number of intersections and numerical integrals needs to be computed in (6.6). Conditioned on $\{E = e\}$, the point process $\{X_i, i = 1, \ldots, Z_X\}$ of thunderstorm cell centers is a Matérn cluster process, which can (approximately) be simulated in W according to Algorithm 3.3.4. A corresponding realization of the (conditional) germ-grain model M is obtained by assigning a disc with radius r to each point of the generated cluster process, see Algorithm 3.3.6. Based on a sequence $M_1, \ldots, M_j$ of $j \in \mathbb{N}$ (conditionally) independent and identically distributed germ-grain models with $M_i \overset{\mathrm{d}}{=} M$ for $i = 1, \ldots, j$ (conditioned on $\{E = e\}$), an estimator $\hat{\pi}(B)$ of $\pi(B)$ is given by

$$
\hat{\pi}(B) = \frac{1}{j}\, \#\{i \in \{1, \ldots, j\} : B \cap M_i \neq \emptyset\}. \tag{6.7}
$$

Both the representation formula (6.6) and the Monte Carlo estimator (6.7) can also be used for the derivation of point probabilities by setting $B = \{t\}$ for any $t \in W$, compare to (6.5).

In Figure 6.8, typical realizations of the germ-grain model M are illustrated for two sample forecast periods. A comparison to the thunderstorm records from NowCastMIX depicted in Figure 6.2 shows that a representation of thunderstorm cells in clusters is much more realistic than the application of the Cox germ-grain model of precipitation cells introduced in Chapter 4, see the realizations in Figure 6.3. However, we still observe significant differences between recorded and generated thunderstorm clusters as



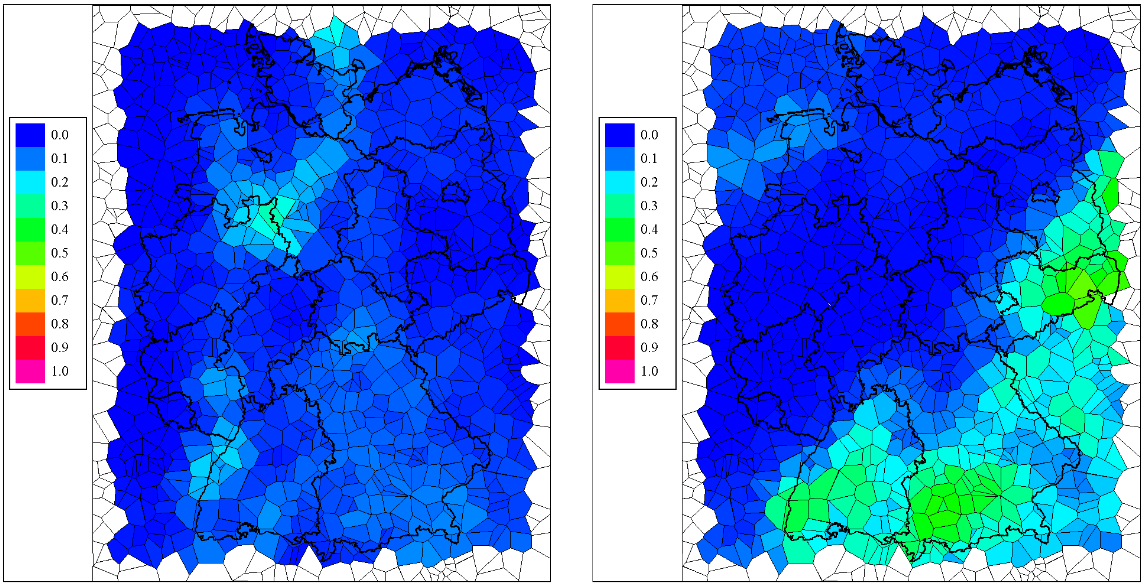
May 22, 2016, 19-20 UTC, $r^{(1)} = 17$ km   July 11, 2016, 14-15 UTC, $r^{(1)} = 25$ km

Figure 6.8.: Realizations of the cluster-based germ-grain model M for the representation of thunderstorm cells with a fixed cell radius of 10 km for two selected forecast periods.

only circular clusters with spatially constant radii and cluster intensities are simulated according to the proposed model. Figure 6.9 shows estimated area probabilities for the test areas $B_1, \ldots, B_{999}$ introduced in Section 6.2 for the same two forecast periods. Area probabilities closely correspond to the underlying point forecasts, see Figure 6.1, but forecasters might wish a better precision and sharpness of estimated probabilities, i.e., high probabilities for areas in which thunderstorms occurred and low probabilities in areas without thunderstorms instead of medium probabilities in large regions, compare to Figure 6.2.

6.5.2. Conditional simulation using thunderstorm records

In order to provide even more realistic realizations of thunderstorm clusters and to further increase forecast quality, we propose a conditional simulation algorithm for the cluster-based germ-grain model M, which generates thunderstorm cells with respect to data from past periods. Let T again denote the one-hour time period for which the forecast is made and $T - d$ with $d \geq 60$ minutes the most recent one-hour period for which recorded thunderstorms from NowCastMIX are available. With $x_1, \ldots, x_p \in W$ for $p \in \mathbb{N}$ we denote the centers of all thunderstorm cells included in NowCastMIX for



May 22, 2016, 19-20 UTC   July 11, 2016, 14-15 UTC

Figure 6.9.: Area probabilities for test areas $B_1, \ldots, B_{999}$ estimated based on repeated simulation of the germ-grain model M of thunderstorm cells for two selected forecast periods. Area probabilities are not shown for areas at the boundaries of the considered domain.

period $T - d$ (if no thunderstorms were recorded, then no conditional simulation is possible). The idea behind the algorithm is that for short forecast ranges (to be more precise, for small d) there is a certain probability that some of these cells still exist (with changed positions) in period T and should thus influence the provided forecast. In order to determine such probabilities, the distribution of lifetimes of thunderstorm cells has been estimated by DWD for thunderstorms with different hail flags, see Figure 6.10 and Wapler (2017). At first, for each $i \in \{1, \ldots, p\}$, the total lifetime of cell x_i is randomly generated based on the estimated distributions (in dependence of the storm's hail flag). Then, the remaining lifetime of cell x_i (from the time it was recorded until its death) is simulated by multiplying the total lifetime with a realization of a standard uniformly distributed random variable. Knowing the exact time when x_i was recorded, we can now easily determine whether the thunderstorm cell x_i still exists during the forecast period T or not. Let $\{\tilde{x}_1, \ldots, \tilde{x}_q\} \subset \{x_1, \ldots, x_p\}$ with $q \leq p$ denote all thunderstorm cell centers from period $T - d$ that still exist in period T. As $\tilde{x}_1, \ldots, \tilde{x}_q$ represent the locations in the interval $T - d$, the random movements of these cells have to be simulated next. Since for each $i \in \{1, \ldots, q\}$, NowCastMIX also provides the movement speed and movement direction of the i-th surviving thunderstorm cell $\tilde{x}_i$ at the time it was recorded, see Section 6.1, we can

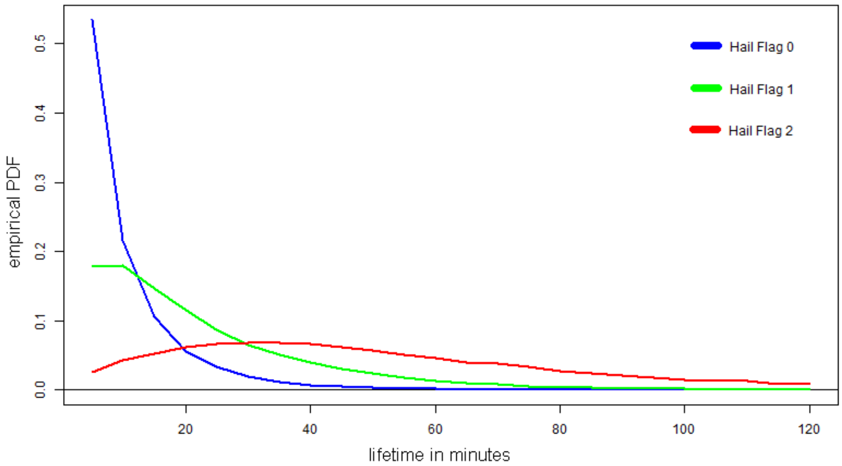


Figure 6.10.: Distribution of the lifetimes of thunderstorm cells for different hail flags.

determine the area of all possible locations $\tilde{x}_i$ can have between the beginning of period T and its death (similar to the warning cones illustrated in Figure 6.2). This area is computed using a propagation angle of 7.5°, making it a triangle or a trapezoid within which we uniformly generate the new location $y_i \in W$ of the i-th surviving cell $\tilde{x}_i$. Thus, $\{y_1, \dots, y_q\}$ can be interpreted as a possible set of thunderstorm cell centers that were recorded in period $T - d$ (at locations $\tilde{x}_1, \dots, \tilde{x}_q$) and still exist in forecast period T.

We propose the following algorithm to generate a realization of the germ-grain model M given the surviving cell centers $y_1, \dots, y_q$.

1. Let $U = \bigcup_{i=1}^{q} b(y_i, r^{(1)})$. Compute the expected numbers $\lambda_{in} = \int_U \lambda^{(0)}(t)\,\mathrm{d}t$ and $\lambda_{out} = \int_{W \setminus U} \lambda^{(0)}(t)\,\mathrm{d}t$ of cluster centers inside and outside of U, respectively.

2. Generate realizations x_{in} and x_{out} of Poisson distributed random variables with parameters λ_{in} and λ_{out}.

3. If possible, simulate x_{in} cluster centers inside U independently according to the intensity function $\{\lambda^{(0)}(t), t \in U\}$ under the condition that each of the discs $b(y_1, r^{(1)}), \dots, b(y_q, r^{(1)})$ contains at least one cluster center. If this is not possible (i.e., if x_{in} is too small), then increase x_{in} by 1 but reduce x_{out} by 1 accordingly and repeat step 3.

4. Simulate x_{out} cluster centers in $W \setminus U$ independently according to the intensity function $\{\lambda^{(0)}(t), t \in W \setminus U\}$. If $x_{out} \leq 0$ due to a possible reduction in step 3, then skip this step and go to step 5.

5. Put a disc with radius $r^{(1)}$ around each cluster center generated in steps 3 and 4 in order to provide the cluster discs.

6. Generate a realization x' of a Poisson distributed random variable with parameter $(x_{in} + x_{out})\, a^{(1)}\pi(r^{(1)})^2$. Put $x = x' - q$, which can be interpreted as the number of thunderstorm cell centers to be simulated.

7. Repeat x times the following. Choose one cluster generated in step 3 or 4 at random and generate a uniformly distributed thunderstorm cell center in the corresponding cluster disc.

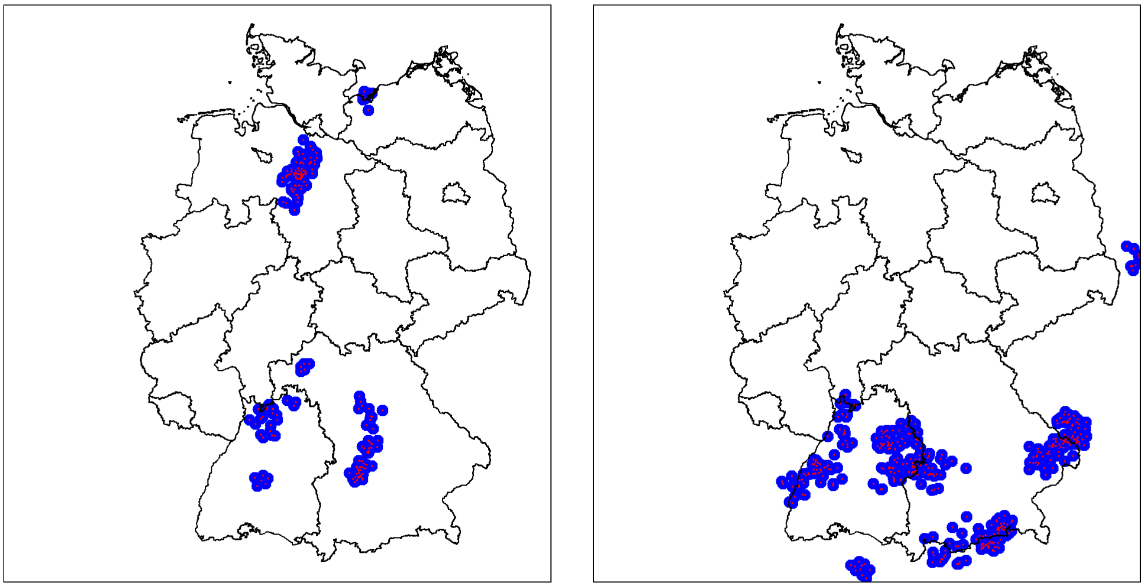
8. Put a disc with radius r around each point generated in step 7 and around each remaining cell center y_i for $i = 1, \ldots, q$. The union of all these discs can be interpreted as a realization of M under the conditions that

   (i) the realization contains the thunderstorm cells with centers $y_1, \ldots, y_q$ and

   (ii) the expected number of generated thunderstorm cells is not changed compared to unconditional simulation.

Note that according to the properties of Poisson point processes, the number of cluster centers $\#\{i : Y_i \in U\}$ in U and the number of cluster centers $\#\{i : Y_i \in W \setminus U\}$ outside U should be independent random variables, see Definition 3.3.9. However, if in step 3 x_{out} is not reduced accordingly if x_{in} is increased to get one cluster center in each disc $b(y_i, r^{(1)})$ for $i = 1, \ldots, q$, too many clusters are generated on average, which introduces a significant model bias in applications. Point and area probabilities for the occurrence of thunderstorms can again be determined using the estimator (6.7), where the underlying realizations of M are generated using the algorithm described above.

We illustrate the value of using the proposed conditional simulation algorithm by showing some example results for the two sample forecast periods considered throughout this chapter. In Figure 6.11, typical realizations of the germ-grain model M of thunderstorm cells are shown that were generated by the conditional simulation algorithm using thunderstorm records of the preceding one-hour time period (i.e., the best case $d = 60$ minutes is considered). We observe that the obtained realizations look much more realistic compared to the ones generated by unconditional simulation, see Figure 6.8. Realized thunderstorm clusters have more general shapes and we observe a high similarity to the corresponding thunderstorm records of NowCastMIX data, compare to Figure 6.2. On the other hand, different realizations still provide a sufficient amount of random variation to account for uncertainties in thunderstorm activity in forecast period T. Finally, Figures 6.12 and 6.13 illustrate examples of point probabilities for the 1,575 points of the 20 km $\times$ 20 km lattice, see Section 6.1, and area probabilities for the test areas $B_1, \ldots, B_{999}$, see Section 6.2, that are estimated using the proposed conditional simulation algorithm, where again the optimal case $d = 60$ minutes is considered. We observe large differences of estimated probabilities



May 22, 2016, 19-20 UTC, $r^{(1)} = 17$ km   July 11, 2016, 14-15 UTC, $r^{(1)} = 25$ km

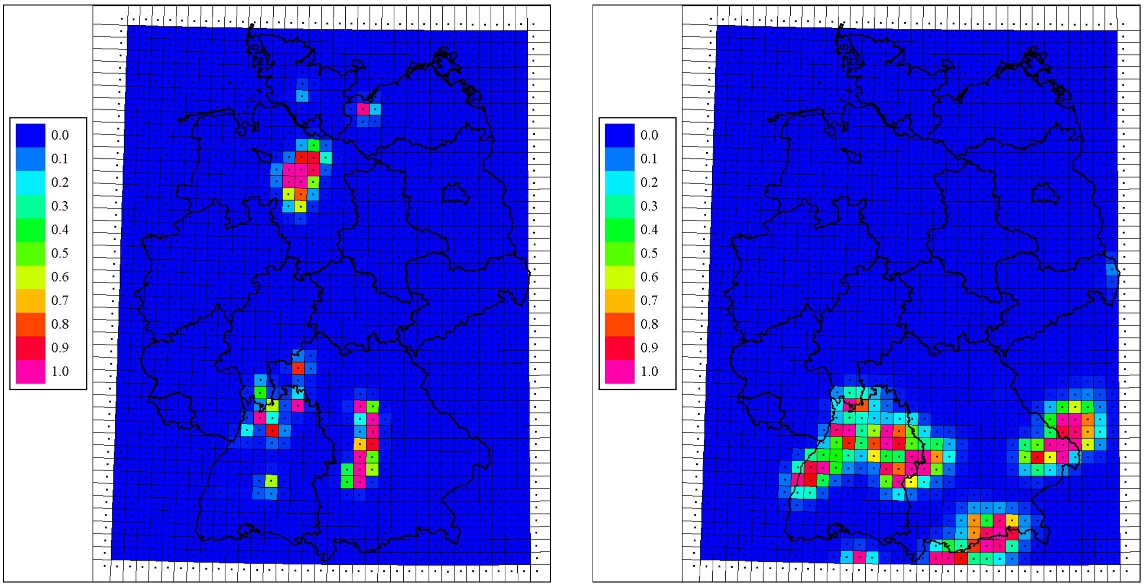
Figure 6.11.: Realizations of the cluster-based germ-grain model M for the representation of thunderstorm cells with a fixed cell radius of 10 km generated by conditional simulation using NowCastMIX data for two selected forecast periods.

to the point probabilities provided by DWD, compare to Figure 6.1, and to the area probabilities estimated based on unconditional simulation, see Figure 6.9. In particular, the sharpness of obtained forecasts is significantly improved and estimated probabilities correspond particularly well to the thunderstorm records obtained from NowCastMIX.

The entire procedure described in this section can as well be applied using not only recorded thunderstorm cells from period $T - d$ but also from earlier periods (depending on how big d is) since thunderstorms, particularly those with hail flag 2, have a good chance to exist two hours or longer. Furthermore, a slightly larger value of d will be considered in most applications for practical reasons.

6.6. Forecast verification

In the final section of this chapter, we perform a verification of point and area probabilities that were computed or estimated using the proposed cluster-based germ-grain model for the representation of thunderstorm cells. For this purpose, we again consider



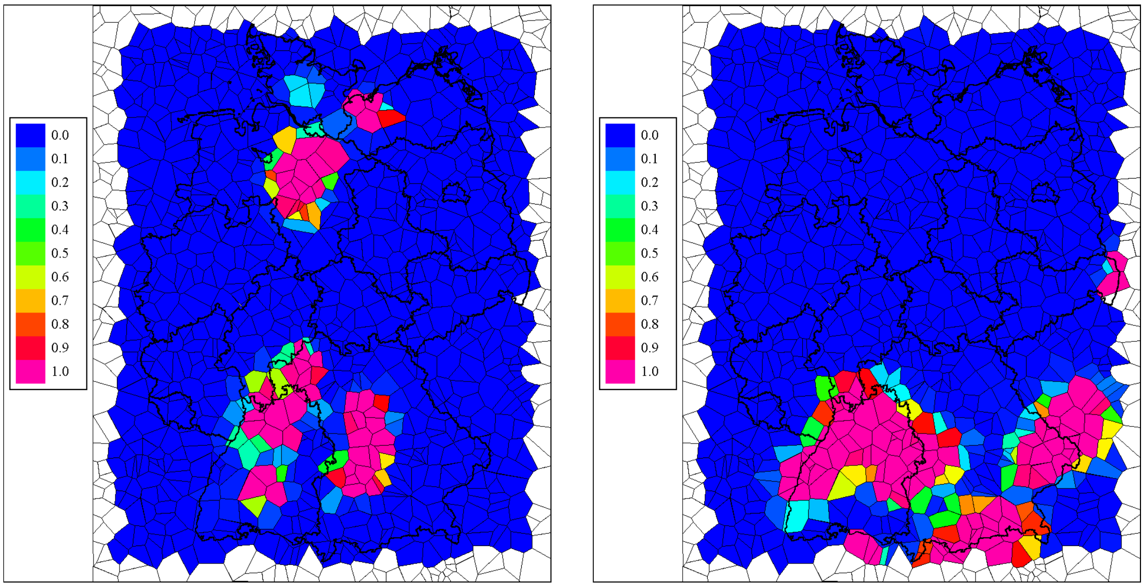
May 22, 2016, 19-20 UTC

July 11, 2016, 14-15 UTC

Figure 6.12.: Point probabilities for the 1,575 points of the 20 km × 20 km lattice estimated based on conditional simulation of the germ-grain model M of thunderstorm cells using NowCastMIX data for two selected forecast periods. Point probabilities are not shown for points at the boundaries of the considered domain.

the forecast periods, point probabilities, test areas, and thunderstorm records from NowCastMIX data that were introduced in Sections 6.1 and 6.2. For the computation of the cluster characteristics $a^{(1)}$ and $r^{(1)}$ as described in Section 6.4.2 and for the application of the conditional simulation algorithm introduced in Section 6.5.2, we assume that forecasts are always made directly at the beginning of the corresponding forecast period and that NowCastMIX data are available for the preceding one-hour period, which means that $d = 60$ minutes. When providing forecasts based on the conditional simulation algorithm, NowCastMIX data is used for the condition and for forecast verification. However, for verification data is considered that corresponds to the respective forecast period, whereas for the condition only data from preceding periods is taken into account.

We again consider the biases, LSSs, and empirical correlation coefficients of estimated probabilities and thunderstorm indicators as defined in Sections 4.7.1 and 6.2. In Figure 6.14, the test areas $B_1, \ldots, B_{999}$ are colored according to the corresponding scores for area probabilities computed by means of the representation formula (6.6). The results reveal that the cluster-based germ-grain model M is much more suitable for the modeling of thunderstorm cells than the Cox germ-grain model considered in



May 22, 2016, 19-20 UTC

July 11, 2016, 14-15 UTC

Figure 6.13.: Area probabilities for test areas $B_1, \dots, B_{999}$ estimated based on conditional simulation of the germ-grain model M of thunderstorm cells using NowCastMIX data for two selected forecast periods. Area probabilities are not shown for areas at the boundaries of the considered domain.

Section 6.2. Compared to Figure 6.4 (right), biases were reduced drastically (now the mean bias is at $-0.4\,\%$, single values are between $-3.5\,\%$ and $2\,\%$). The lowest and highest biases occur in regions, where similar biases are also present in the underlying point probabilities, see Figure 6.4 (left). LSSs are positive for all test areas over land (mean value of 0.28, single values ranging between 0.15 and 0.45) and thus clearly higher than for the Cox germ-grain model, see Figure 6.5 (right). Correlation coefficients have similar values as for the Cox germ-grain model (mean value of 0.4, most single values between 0.25 and 0.6), compare to Figure 6.6 (right).

Finally, we analyze the performance of point and area probabilities that are estimated using the conditional simulation algorithm proposed in Section 6.5.2. In Figures 6.15 and 6.16, the three considered score functions are illustrated for point and area probabilities. When comparing estimated point probabilities with those probabilities from the available data, see Figures 6.4-6.6 (left), we find that LSSs (mean value 0.51, single values reaching up to 0.7) and empirical correlation coefficients (mean value 0.63, single values reaching up to 0.8) are considerably higher when using conditional simulation. Furthermore, no model bias is introduced. The results for area probabilities are even more convincing. A comparison of scores with those in Figure 6.14 shows drastically increased LSSs (mean value of 0.49, single values up to 0.65) and empirical correlation coefficients (mean value of 0.71, almost all single values between 0.55 and 0.8) together with a mean bias of less than 1 %. This shows impressively that using the proposed conditional simulation algorithm for the cluster-based germ-grain model of thunderstorm cells does not only provide realistic realizations of thunderstorm events (see Figure 6.11) but also allows for an estimation of both point and area probabilities with a very high precision. Thus, we conclude that the models and methods developed in the present chapter provide a valuable tool in the forecasting of thunderstorm events for short forecast ranges that can be applied in operational weather prediction.

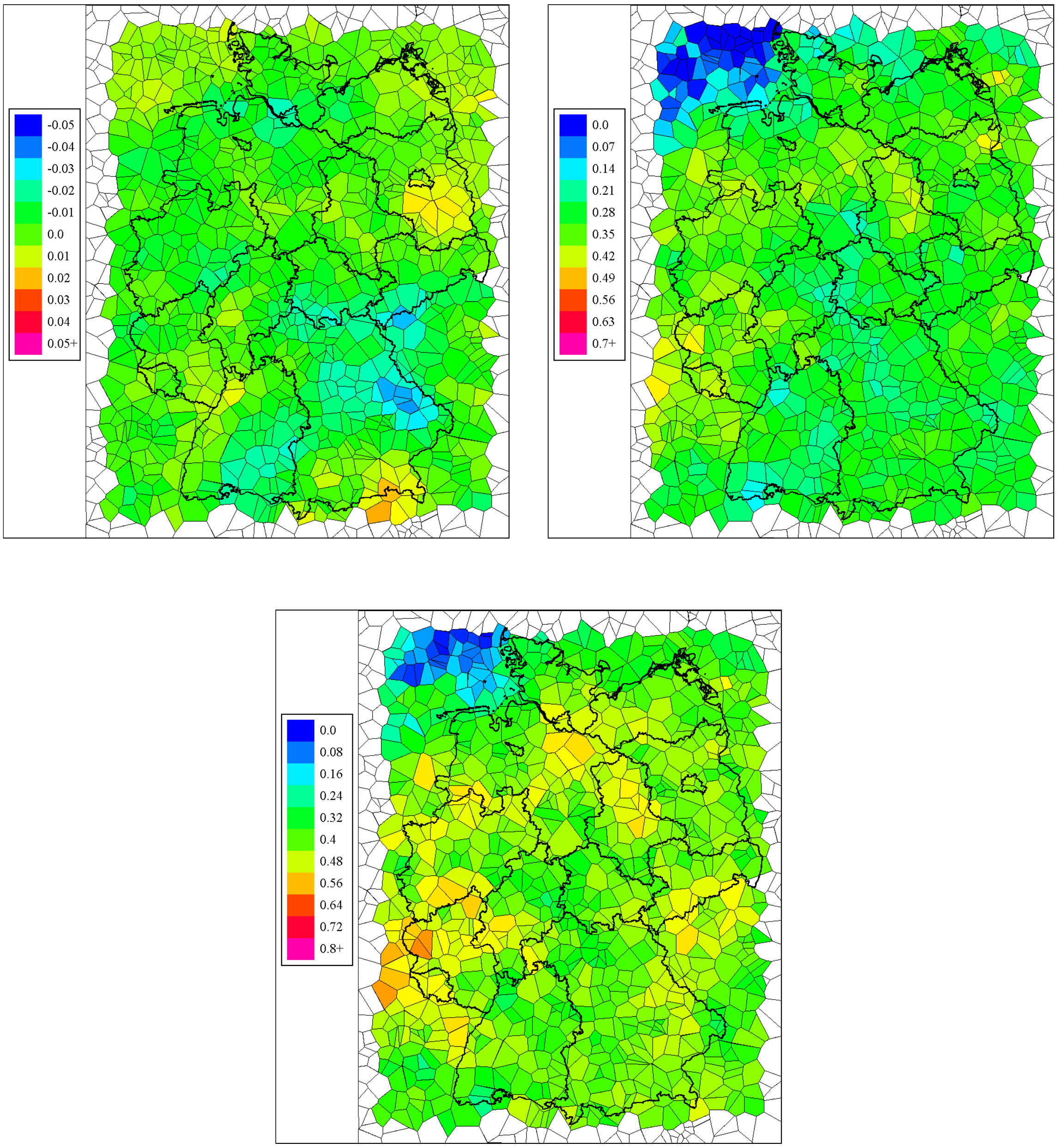


Figure 6.14.: Comparison of area probabilities for test areas $B_1, \ldots, B_{999}$ computed according to the cluster-based germ-grain model M of thunderstorm cells with thunderstorm indicators from NowCastMIX data: biases (top left), LSSs (top right), and empirical correlation coefficients (bottom). Scores are not shown for areas at the boundaries of the considered domain.

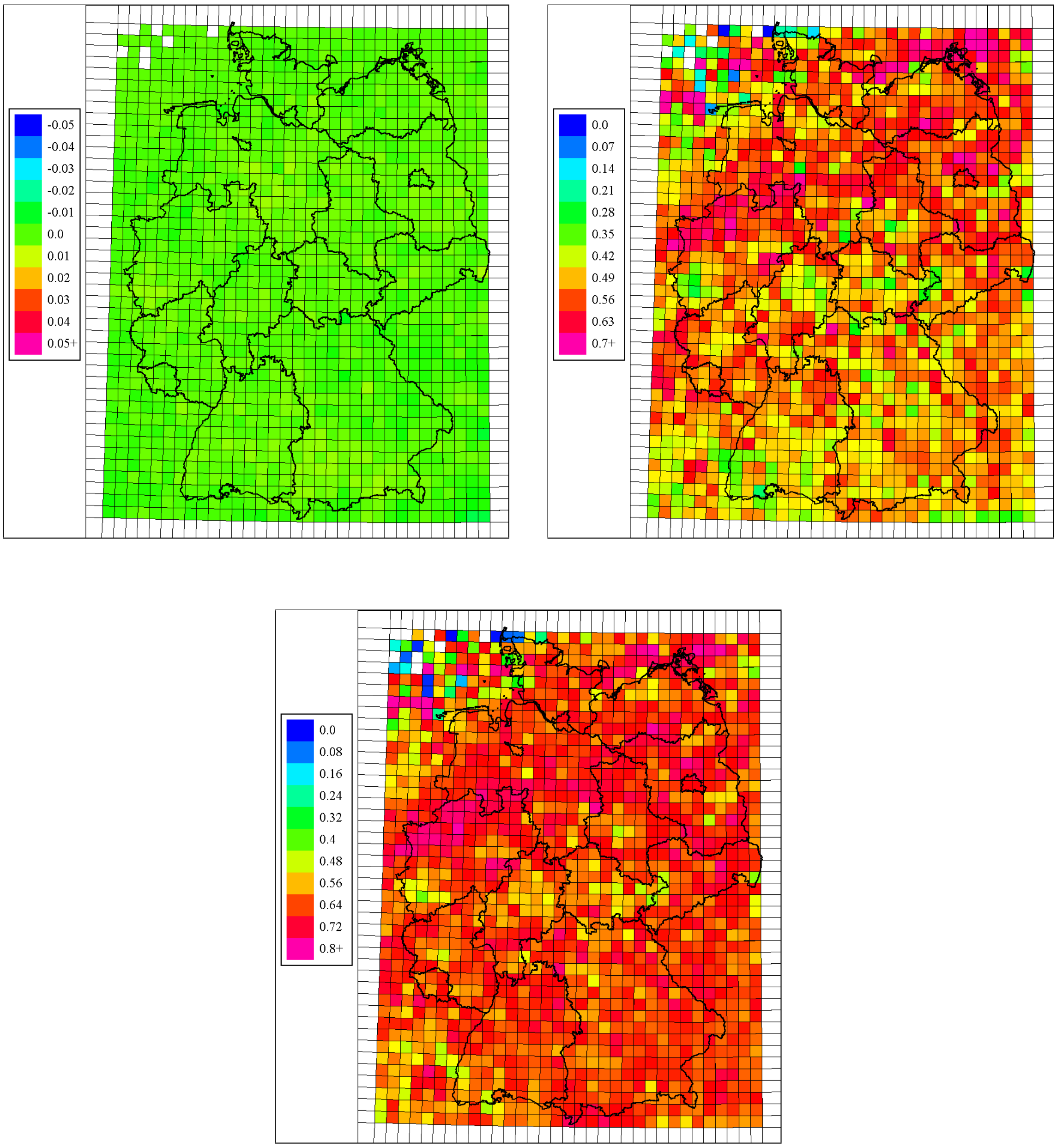


Figure 6.15.: Comparison of point probabilities for the 1,575 points of the 20 km×20 km lattice estimated based on conditional simulation of the cluster-based germ-grain model M of thunderstorm cells with thunderstorm indicators from NowCastMIX data: biases (top left), LSSs (top right), and empirical correlation coefficients (bottom). Scores are not shown for points at the boundaries of the considered domain.

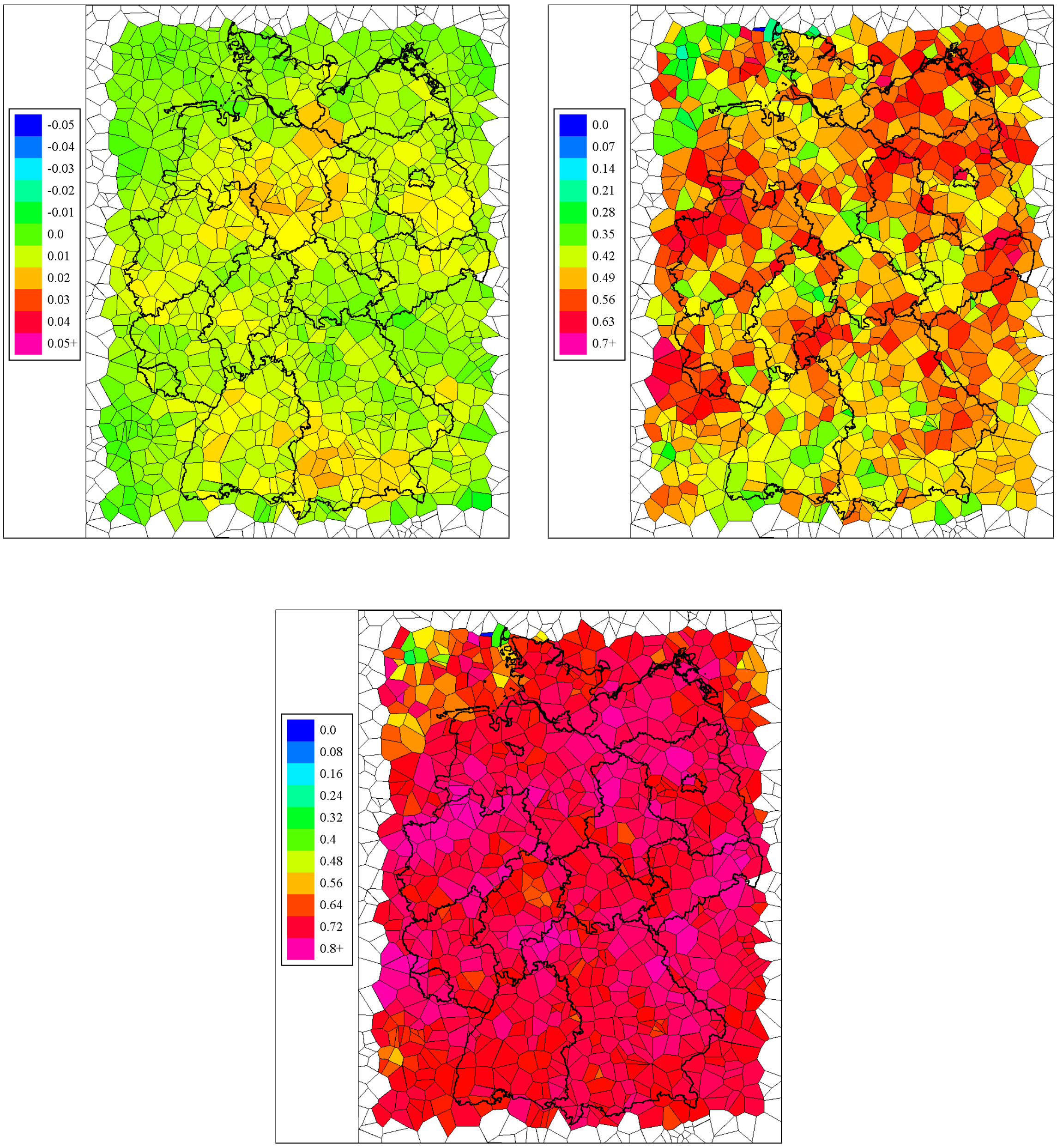


Figure 6.16.: Comparison of area probabilities for test areas $B_1, \ldots, B_{999}$ estimated based on conditional simulation of the cluster-based germ-grain model M of thunderstorm cells with thunderstorm indicators from NowCastMIX data: biases (top left), LSSs (top right), and empirical correlation coefficients (bottom). Scores are not shown for areas at the boundaries of the considered domain.

Part III.

Statistical analysis of paleogeographical space-time data

7. Spatio-temporal distribution of prehistoric populations in North America

In Part III of the present thesis, we perform a qualitative statistical analysis that aims to find systematic relationships and interactions between prehistoric human populations and vegetation composition in North America during the Holocene. However, for that purpose, spatio-temporal estimates of population and vegetation intensities need to be determined first using comprehensive archaeological and paleoenvironmental databases. In this chapter, we propose a statistical approach to the estimation of spatially smoothed population intensity maps for a sequence of time intervals covering the past 13000 years. In particular, this involves the use of nonparametric kernel methods, which have successfully been applied to archaeological data before, see, e.g., Collard et al. (2010) or Grove (2011). In Section 7.1, we present the underlying data, which are extracted from an extensive database of archaeological samples from North America. These data are used in Section 7.2 to estimate prehistoric population intensity maps, where a particular focus is put on various potential errors such as taphonomic loss, inhomogeneous sampling strategies, and boundary effects. In Section 7.3, we show some example maps and give a brief summary of results. Finally, Section 7.4 provides a sensitivity and robustness analysis by modifying the applied nonparametric kernel methods and the underlying database to demonstrate the significance of obtained population patterns. The results presented in this chapter were summarized in Chaput et al. (2015) (with a focus on discussion of estimated population maps). A similar study has recently been performed for South America, see Goldberg et al. (2016).

7.1. Description of data

In order to reconstruct and analyze paleodemographic changes in North America during the Holocene, comprehensive archaeological databases are frequently used. We follow the generally acknowledged assumption that the frequency of dated archaeological samples is, after accounting for several potential biases, proportional to population density, which is denoted as ‘dates as data’ approach in literature, see, e.g., Steele (2010).

In the present chapter, we use data from the Canadian Archaeological Radiocarbon Database (CARD), see Gajewski et al. (2011), which was created by Dr. Richard Morlan of the Canadian Museum of History and which is now maintained and regularly updated jointly by the Canadian Museum of History and the Laboratory of Archaeology at the University of British Columbia. The version of the CARD considered in this thesis contains 35,905 samples collected from 9,149 geographically distinct sites in Canada, the United States (US), and eastern Russia. Currently, great effort is made to extend CARD to a global database by adding samples from all continents (except the Antarctic) but in this chapter, analysis is restricted to North America.

Each entry in the CARD represents an archaeological sample, which is characterized by a variety of information. Most importantly, the geographical coordinates of the archaeological excavation site where the sample was found are recorded including the elevation and the corresponding province (for sites in Canada) or state (for sites in the US). Of similar relevance is the age of the sample, which is estimated using the radiocarbon method, see Libby (1955). This method relies on the observation that radiocarbon (^{14}C, a radioactive isotope of carbon with 8 neutrons) is constantly created in the earth's atmosphere and absorbed by plants, animals, and humans, leading to a relatively stable amount of ^{14}C in all living organisms. Upon death, this absorption stops and the radioactive ^{14}C within the organism starts to decay according to a well-known half-life period. Analyzing the amount of ^{14}C in a sample of a dead plant, animal or human thus allows to determine the age of the sample. However, the radiocarbon method falsely assumes that the ratio of ^{14}C and ^{12}C (non-radioactive isotope of carbon with 6 neutrons) is constant over time, which is why derived (uncalibrated) radiocarbon ages are subject to a systematic bias. To account for that, an adjustment is made using the IntCal04 calibration curve, see Reimer et al. (2004), where an entire probability distribution is determined for the (unknown) age of each sample in the CARD. The median of this distribution is then used as a calibrated estimate (denoted as calibrated age or age in calender years) of the sample's age. At the time this thesis was written, an updated version, IntCal13, was available, see Reimer et al. (2013), but the differences to IntCal04 are marginal for radiocarbon ages of less than 15000 years. Obtained calibrated ages are typically provided in years before present (BP), where by convention, the term 'present' refers to the year CE 1950 (e.g., the year 600 BP is equivalent to the year CE 1350). Besides geographical locations and calibrated ages, each sample in the CARD includes a variety of additional information such as classification and type of the dated material, archaeological provenance, site names and components, references, contributing collectors, collection dates, and general comments. However, the only further attribute considered in this thesis is a classification to distinguish between cultural (i.e., related to human activity) and paleoenvironmental (i.e., related to animals and plants) dates. Unfortunately, one has to notice that most attributes are not provided for all samples. For example, locations are available for 99.8 %, calibrated ages for 94.03 %, and classification for about 99.9 % of dates in the CARD. Furthermore,

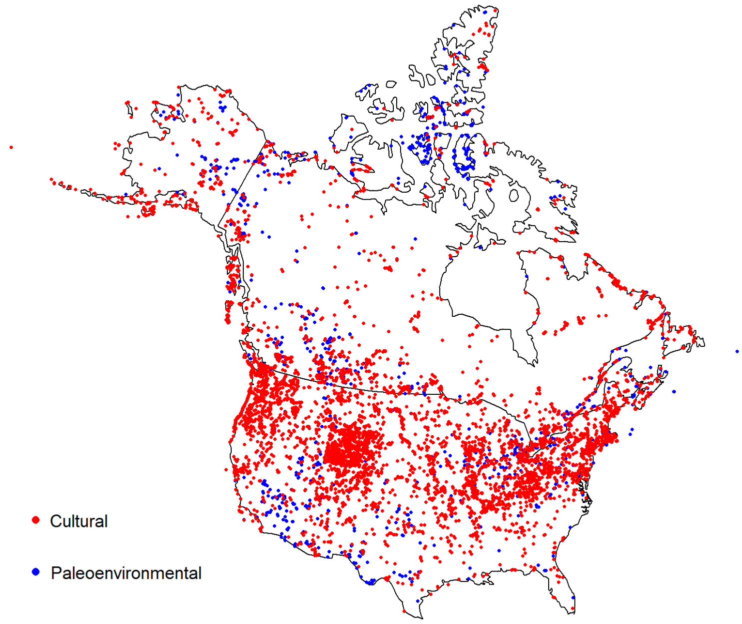


Figure 7.1.: Sites with cultural and paleoenvironmental dates from the CARD.

dummy values such as ‘unknown’ occasionally occur and need to be identified and eliminated.

Altogether, we use 33,696 samples (77 are missing locations, 61 are from Russia and 2,071 are missing radiocarbon ages or classification) from 7,754 geographically distinct sites, where a majority of 29,609 dates is classified as cultural and 4,087 dates have a paleoenvironmental context, see Figure 7.1. Especially for cultural dates we observe large differences in local sampling densities. Plenty of samples were obtained in the northeastern US and regions that roughly correspond to the states of Wyoming, Utah, Colorado, Washington, and Oregon. The south of the US as well as central and northern Canada, however, seem to be underrepresented with very low sampling densities. Possible reasons causing these inhomogeneities are that in areas of higher sample densities often more research institutions with interest in archaeological field work are located and that archaeologists tend to investigate regions where evidence of prehistoric population has already been found (i.e., where sampling density is already high). A proper adjustment for inhomogeneous sampling strategies in the context of estimating population intensity maps is proposed in Section 7.2.

To conclude this section, we analyze the temporal distribution of the 29,609 cultural dates from the CARD, see the histogram in Figure 7.2. Apparently, there is a large amount of dates with calibrated ages between 500 BP and 1500 BP (about 40 % of all cultural samples). With increasing age the number of available dates quickly decreases,

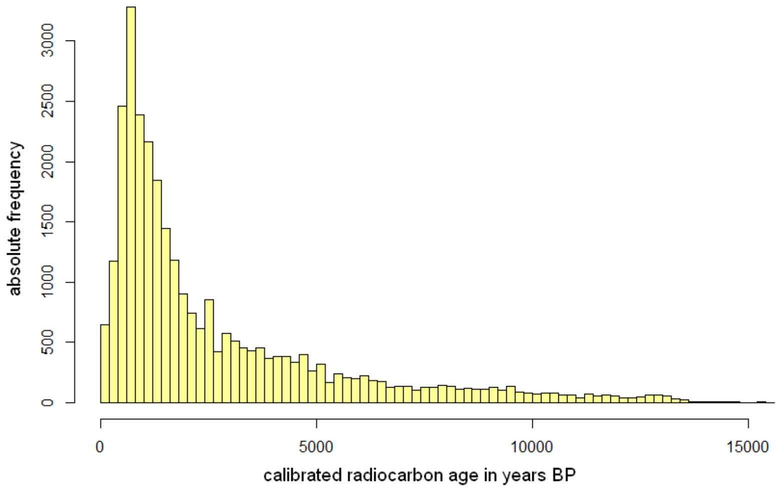


Figure 7.2.: Temporal distribution of calibrated radiocarbon ages for cultural dates from the CARD.

which is most likely caused by both the destruction of older carbon due to erosion and dissolution (taphonomic loss) and lower population numbers for earlier time periods. There are very few dates found with ages older than 13000 BP, which corresponds well with the prevailing opinion that a settlement of the North American continent started only a couple of thousand years before that time. We also identified some dates having ages of 20000 years or more, which are, however, considered to be subject to dating errors or misspecification. Surprisingly, prior to 800 BP the frequency of dates also decreases showing considerably smaller numbers for the most recent 500 years. A potential reason for this is that archaeological samples from the era after the arrival of Europeans can often be identified and dated without using the radiocarbon method, making such dates occur less frequently in the CARD. However, it also seems possible that this effect indeed corresponds to a decline in population numbers, which could be due to the spread of European diseases, see O'Fallon and Fehren-Schmitz (2011).

7.2. Nonparametric estimation of population intensity maps

In this section, we propose a statistical method to the estimation of spatially smoothed population intensity maps based on the locations and calibrated radiocarbon ages of archaeological samples from the CARD. The underlying assumption of this approach is that on the continental scale of this study, a greater number of dates is indicative of a higher population activity ('dates as data approach') when potential biases such as inhomogeneous sampling strategies and taphonomic loss are accounted for appropriately. Since the development of spatial or spatio-temporal parametric models for population intensities seems to be a highly complicated task, we propose the application of nonparametric kernel methods. To allow for a comparison of estimated population intensity maps over time, we consider a sequence of 121 500-year intervals, which range from $500 - 1000$ BP, $600 - 1100$ BP, $\ldots, 12500 - 13000$ BP. An interval length of 500 years is selected as a compromise to avoid aggregating too many dates and having too few to produce reliable estimates later on. For a fixed 500-year interval, denoted as $I = [y - 250, y + 250]$ with $y \in \{750, 850, \ldots, 12750\}$ in the following, we define the compact set $W \subset \mathbb{R}^2$ describing all locations of North America (including some bigger offshore islands) that are not covered by ice in year y (reconstructions of prehistoric ice sheets were used to identify glaciated areas in the past 13000 years, see Dyke, 2002). In particular, each location $t \in W$ can be represented as $t = (t^{(1)}, t^{(2)})$ with $t^{(1)}$ and $t^{(2)}$ being the longitude and latitude of t in degree, i.e., W is a geographical domain, see Remark 3.2.4. By $\{\lambda(t), t \in W\}$ we denote the unknown deterministic population intensity map for interval I. We suggest to estimate $\{\lambda(t), t \in W\}$ as follows. At first, we select all cultural dates from the CARD with calibrated ages in the extended time interval $\tilde{I} = [y-650, y+650]$ (that starts 400 years earlier and ends 400 years later than I) and denote by $t_1, \ldots, t_n \in W$ with $n \leq 29,609$ the locations of the archaeological sites at which the corresponding samples were excavated. In most time intervals, some of the sites $t_1, \ldots, t_n$ are identical since different samples can be found at the same archaeological site. Dates within 400 years before or following the fixed time interval I are also taken into consideration because they can still have an influence on population intensities in I (e.g., the number of people at a given time and location is influenced by how many people were there several generations before) and to account for possible errors occurring from radiocarbon dating and calibration. However, dates with ages in $\tilde{I} \setminus I$ should have a smaller impact than those with ages in I, which is why a weight w_j is assigned to site t_j for each $j \in \{1, \ldots, n\}$. If the calibrated age of the date that corresponds to site t_j falls into I, then we set $w_j = 1$, otherwise $w_j < 1$ decreases linearly with the temporal distance of the age to the boundary of I. In particular, for dates with an age of $y - 650$ BP or $y + 650$ BP we get that $w_j = 0$.

We suggest to use a modified two-dimensional KDE, see Section 3.4.2, for the estimation of $\{\lambda(t), t \in W\}$. An intuitive idea is to put a disc $b(t_j, h)$ with a radius $h > 0$ in km around the site t_j for each $j \in \{1, \ldots, n\}$ in order to describe an area where human population could have occurred in the interval I based on the archaeological sample found at site t_j (note that all discs are computed using the great-circle distance in this chapter, see Remark 3.2.4). The radius h is chosen to be globally fixed as all cultural dates are expected to indicate population activity within the same range of the corresponding excavation sites. In particular, we ignore geographical features such as rivers or mountains, which seem impossible to be modeled in a continent-wide analysis. Next, we account for the fact that the cultural date found at t_j is more likely to indicate population activity close to t_j than near the boundary of $b(t_j, h)$. For that purpose, a (scaled) radially symmetric Epanechnikov kernel $\kappa_h^E(\cdot, t_j)$ with bandwidth matrix $H = \mathrm{diag}(h^2, h^2)$ is assigned to each disc $b(t_j, h)$, where $\kappa_h^E : W \times W \to [0, \infty)$ is given by

$$\kappa_h^E(s,t) = \frac{1}{h^2}\left(1 - \frac{d_{GC}(s,t)^2}{h^2}\right)\mathbb{1}_{[0,h]}(d_{GC}(s,t)), \quad s,t \in W, \tag{7.1}$$

compare to Section 3.4.1 and (3.33). Since W describes a geographical domain, i.e., all locations in W are represented by geographical coordinates, the great-circle distance is used instead of the Euclidean distance in (7.1), see (3.4) and Remark 3.4.3. As a consequence, the Epanechnikov kernels $\kappa_h^E(\cdot, t_1), \ldots, \kappa_h^E(\cdot, t_n)$ do not necessarily integrate to 1 over W, which is why they are denoted as scaled kernels here.

Generally speaking, the considered nonparametric approach with a globally fixed bandwidth h suggests that a cultural date from the CARD is indicative for a certain population number, which is constant for all dates. A problem occurs if a site t_j is located close to the coast, a region permanently covered by ice or the boundary to Central America such that the disc $b(t_j, h)$ intersects with an area where no population is possible or no data are available. This, however, would cause such dates to have less influence on population intensities in the domain W, which leads to estimated population intensity maps that decrease towards the boundaries of W. To account for this, a boundary correction is proposed, where for each $j \in \{1, \ldots, n\}$ the integral

$$e_j^{(h)} = \int_W \kappa_h^E(t, t_j)\,\mathrm{d}t$$

of the scaled Epanechnikov kernel $\kappa_h^E(\cdot, t_j)$ centered at t_j is calculated numerically and $\kappa_h^E(\cdot, t_j)$ is then multiplied by $1/e_j^{(h)}$. In particular, a site t_j close to the boundary gets a higher scaling factor (due to a smaller integral $e_j^{(h)}$) than a site in the central US. Furthermore, all scaled kernel functions are now guaranteed to integrate to 1.

Another potential error that needs to be addressed is the taphonomic bias inherent in the CARD due to the long-term loss of samples caused by, e.g., erosion and dissolution. In particular, estimates of population intensities are expected to be systematically too low for earlier time intervals due to older dates being underrepresented in the CARD. To correct for taphonomic loss, a temporal loss function is estimated in Surovell et al. (2009) based on ice core and geologic volcanic activity records, which has been successfully applied to the estimation of prehistoric demography, see, e.g., Peros et al. (2010). We assume that taphonomic loss rates are constant for the entire continent, although we are aware that degradation may be slightly lower in the Arctic, where permafrost preserves archaeological materials. Accordingly, we suggest to compute a taphonomic loss corrector c_I for the considered 500-year interval I by applying the loss function of Surovell et al. (2009) to the center y of the interval I, i.e.,

$$c_I = 5.726442 \cdot 10^6 (y + 2176.4)^{-1.3925309}.$$

Combining the scaled kernel functions $\kappa_h^E(\cdot, t_1), \dots, \kappa_h^E(\cdot, t_n)$ with the weights $w_1, \dots, w_n$, the boundary correctors $e_1^{(h)}, \dots, e_n^{(h)}$, and the taphonomic loss corrector c_I provides the following estimate $\{\tilde{\lambda}(t), t \in W\}$ of the population intensity map $\{\lambda(t), t \in W\}$ for the 500-year interval I:

$$\tilde{\lambda}(t) = \frac{1}{c_I} \sum_{j=1}^{n} \frac{w_j}{e_j^{(h)}} \kappa_h^E(t, t_j), \quad t \in W. \tag{7.2}$$

A fundamental question is the choice of a proper bandwidth h as varying h has a significant effect on obtained population intensity maps, see Section 7.4.1. Small values of h often show many local details but could be inappropriate to provide reliable estimates in regions with only few sites available and too large values of h are expected to overgeneralize the resulting patterns of population activity. A frequently used option is to algorithmically determine the bandwidth based on the underlying data, see Section 3.4.4, in which case LCV is the most popular choice. However, LCV is very sensitive to outliers and multiple data points and requires long computation times when large datasets are considered, making it a poor choice in our context. A more applicable alternative is given by Scott's rule, which we use to determine a bandwidth for each of the considered 500-year intervals, see Figure 7.3. In order to provide comparability of population intensity maps over time, we aim to choose a constant bandwidth for all 121 time intervals. For that purpose, we suggest to use $h = 600$ km, which is the rounded mean value of the single bandwidths determined according to Scott's rule. This choice turns out to be a good compromise, see the example maps illustrated in Sections 7.3 and 7.4.1.

Unfortunately, $\{\tilde{\lambda}(t), t \in W\}$ turns out to be a poor estimate of the population intensity map $\{\lambda(t), t \in W\}$ as it does not account for inconsistencies in regional sampling effort. For example, a number of ten dates found in a region with only two excavation sites should indicate a higher population intensity than the same number of dates found in a region with equal size and eight excavation sites. If this remains unaccounted for, estimated population intensities are highly correlated to sampling intensities because a greater sampling effort produces a larger number of dates, which results in higher population estimates. For that purpose, a sampling intensity map is determined using (7.2), where $t_1, \dots, t_n$ are replaced by all considered 7,754 geographically distinct sites $\tilde{t}_1, \dots, \tilde{t}_{7754}$ of the CARD (excluding sites with missing attributes but including those with paleoenvironmental dates) with $w_1, \dots, w_{7754}$ and c_I set to 1 and using similar boundary correctors $\tilde{e}_1^{(h)}, \dots, \tilde{e}_{7754}^{(h)}$ for the Epanechnikov kernels $\kappa_h^E(\cdot, \tilde{t}_1), \dots, \kappa_h^E(\cdot, \tilde{t}_{7754})$ centered at $\tilde{t}_1, \dots, \tilde{t}_{7754}$. In Figure 7.4, the estimated sampling intensity map with

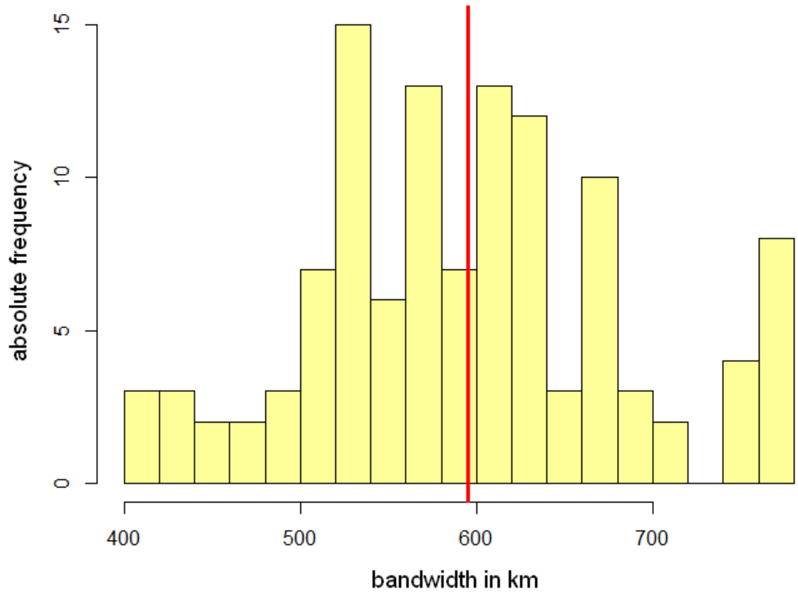


Figure 7.3.: Histogram of bandwidths determined according to Scott's rule for the 121 considered 500-year intervals. The red line indicates the mean bandwidth of 595 km.

bandwidth $h = 600$ km is illustrated, where all intensities are scaled to the interval $[0, 1]$. A comparison to Figure 7.1 shows a close visual correspondence between estimated sampling intensities and the underlying data. Finally, the raw population estimates obtained according to (7.2) are divided by estimated sampling intensities to adjust for inhomogeneous sampling strategies. This results in a more reliable estimate $\{\hat{\lambda}(t), t \in W\}$ of $\{\lambda(t), t \in W\}$ given by

$$\hat{\lambda}(t) = \frac{1}{c_I} \frac{\sum_{j=1}^{n} \frac{w_j}{e_j^{(h)}} \kappa_h^E(t, t_j)}{\sum_{k=1}^{7754} \frac{1}{\tilde{e}_k^{(h)}} \kappa_h^E(t, \tilde{t}_k)}, \quad t \in W, \tag{7.3}$$

which accounts for possible dating and calibration errors, boundary errors, temporal bias due to taphonomic loss, and sampling bias. Heuristically speaking, estimated population intensities can roughly be interpreted as smoothed numbers of dates per site, which should better reflect population intensities than only considering numbers of dates. The bandwidth $h = 600$ km is chosen large enough such that the denominator in (7.3) is positive for each location $t \in W$ and each considered interval I.

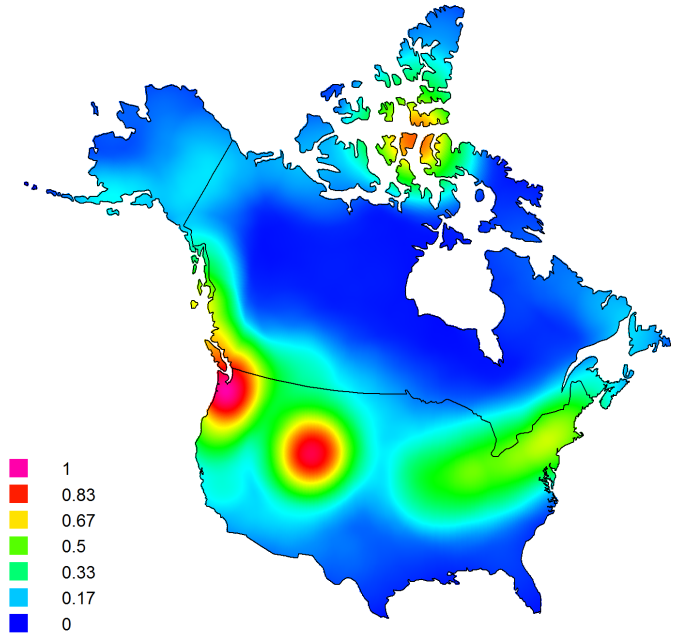


Figure 7.4.: Sampling intensity map estimated based on 7,754 geographically distinct sites from the CARD with a fixed bandwidth of $h = 600$ km.

7.3. Presentation of results

We present and discuss estimated population intensity maps obtained according to (7.3) with a bandwidth of $h = 600$ km. Since we cannot show all results here, Figure 7.5 illustrates estimated maps for three selected 500-year intervals. In order to allow for a comparison of maps through time and to make temporal changes more apparent (especially during earlier intervals), we rescale all estimated intensities to $[0, 1]$ and plot population intensity maps of all time intervals on a joint logarithmic scale. In the following, we briefly describe results for different regions of the North American continent. A much more detailed discussion in the context of archaeological literature is provided in Chaput et al. (2015).

In Alaska, population intensities fluctuated periodically for the entire time period, which is reasonable considering that Alaska was the location of repeated migrations into North America. Population intensities were relatively high in northern and central Alaska between 13000 BP and 8000 BP, see Figure 7.5 (bottom), with a peak at 11500 BP, which is most likely related to Paleoindian populations. Around 6000 BP, a second increase started on the Aleutian Islands and slowly moved east over the next 1500 years, which agrees with archaeological evidence that Aleut peoples colonized these islands at this time. Beginning at 4500 BP, a 1000 year long increase in population intensities at Alaska is observed, which coincides with a suspected migration of Paleoeskimo from

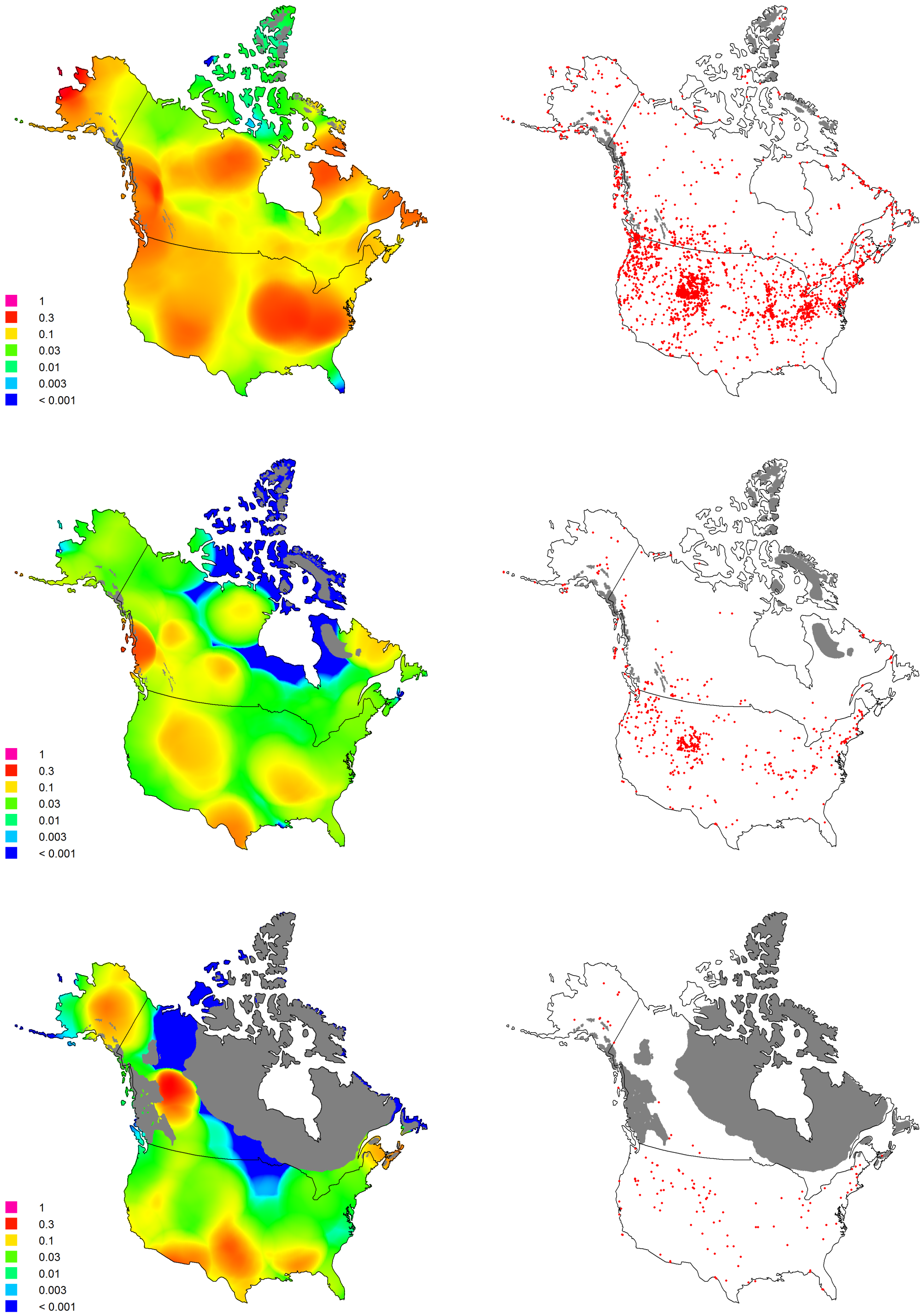


Figure 7.5.: Population intensity maps estimated using a modified KDE with a fixed bandwidth of $h = 600$ km (left) and underlying dates from the CARD (right) for three selected 500-year intervals: 1600-2100 BP (top), 6400-6900 BP (center), and 12100-12600 BP (bottom). Gray colors indicate areas permanently covered by ice.

Siberia. Estimated population intensities show a decrease for some hundred years and then begin to increase at the northern coast (2500 BP) and along the west coast (2000 BP), see Figure 7.5 (top). Finally, at 1500 BP, the majority of Alaska seems occupied, which is most likely associated to the development of Thule culture beginning ca. 1000 BP.

The Canadian Arctic seems to be the last region of the North American continent to be colonized. Prior to 10000 BP, large parts of the Arctic were permanently covered by ice making human population impossible, see Figure 7.5 (bottom). Occupation started around 7000 BP, where the maps show a rapid increase in population intensities in coastal northwestern Canada, including the Mackenzie Delta, Banks Island, and western Victoria Island, see Figure 7.5 (center). Shortly thereafter, populations occurred in the northern Arctic on Ellesmere Island, which supports recent genetic evidence suggesting there was human activity in the general region at this time. By 4000 BP, there were signs of occupation across the entire Canadian Arctic, which remained relatively constant except some fluctuation in the far north. However, population intensities in the Arctic are smaller than in most other regions, see Figure 7.5 (top).

Before and during the Last Glacial Maximum, human migration from Alaska into western Canada via an interior route was not possible due to the presence of ice sheets. At ca. 14000 BP, an ice-free corridor opened between the Laurentide and Cordilleran Ice Sheets and population intensities started to increase rapidly in this region at 13000 BP, see Figure 7.5 (bottom). Intensities remained high until 12000 BP, when the Cordilleran Ice Sheet retreated and westward human expansion into this new terrain probably occurred. Between 11500 BP and 10000 BP, population intensity maps show strong human activity along the entire west coast of British Columbia. After a period with relatively constant population across western Canada, a second increase of estimated intensities at the coast started at 7000 BP, see Figure 7.5 (center), followed by an increase further east. After 5500 BP, population patterns have expanded, potentially due to the development of new trap and tool technologies east of the Rocky Mountains. By 1000 BP, large parts of western Canada were densely populated, which reflects the demographic success of complex hunter–gatherer cultures, see Figure 7.5 (top).

In the eastern US, estimated maps show an increase in population intensities, which began around 10500 BP and expanded northward and westward during the following 2000 years. These results are plausible as numerous archaeological studies indicate a strong presence of Paleoindians in the southeast before 10000 BP. Population intensities decreased east of the Appalachians after 9000 BP and moved towards Missouri, where the presence of Paleoindian populations has been confirmed. Intensities fluctuated until 5000 BP, see Figure 7.5 (center), afterwards populations grew east and west of the Appalachians as well as in the Atlantic and New England regions. After 3500 BP, estimated maps showed greatly increased population activity in the central part of the eastern temperate deciduous forest and after 2000 BP, the highest estimates were in the regions of Ohio and Kentucky, see Figure 7.5 (top). These patterns then persisted until the time Europeans arrived. Altogether, the eastern US were a permanent center of strong human activity during almost all time intervals after 10000 BP.

When analyzing Atlantic Canada, population activity due to Paleoindian peoples was first observed in Nova Scotia and New Brunswick between 13000 BP and 12000 BP, see Figure 7.5 (bottom). Afterwards, estimated population intensities decreased as temperature and dryness increased during the Younger Dryas. Following this, population estimates fluctuated and began to increase around 10000 BP in southeastern Canada and around 8500 BP in Newfoundland and Labrador, where estimates almost tripled. While population intensities decreased again rather quickly in eastern Quebec and central Labrador, they remained stable in the northeast for almost 2000 years. Starting at 5000 BP, estimated intensity maps showed a large growth of population numbers in Atlantic Canada, first in the east and north and several hundred years later in the eastern Hudson Bay region and along the coastline of Labrador. Population intensities decreased in central Quebec but remained high in Newfoundland, Labrador, and the area surrounding New Brunswick until 1500 BP, see Figure 7.5 (top). Some hundred years later, large populations seem to have spread across the entire region.

It remains to consider the central and western US. Population intensity maps for the oldest time periods show two centers of high estimates in Arizona and at the border between Colorado and Kansas, areas where the presence of Paleoindians has been confirmed. Around 12500 BP, these centers moved to Texas, where strong population activity is observed until 5000 BP, see Figure 7.5 (center) and (bottom). By 10500 BP, an intensification of population appeared in California and persisted until 8000 BP, which is likely related to offshore communities using marine resources. The latter half of the Holocene is characterized by an increase of populations in Idaho that moved towards the coast around 4000 BP.

In summary, we find that estimated population intensity maps show clear patterns of population activity, which are considered to capture paleodemographic trends. This is justified by the observation that population patterns correlate well with previous archaeological interpretations of population change across the North American continent during the Holocene that were derived partly from radiocarbon dates, but also other data sources and inferences, see Chaput et al. (2015).

7.4. Sensitivity and robustness analysis

In order to assess the quality and completeness of the CARD as well as the sensitivity of the applied statistical methodology and the robustness of obtained results, we perform some additional analyses using modified estimators and databases.

7.4.1. Modification of estimators

At first, a comparison of population intensity maps obtained according to the modified KDE in (7.3) with different bandwidths between 200 km and 1500 km is made. Figure 7.6 illustrates estimated population intensity maps with $h = 300$ km (left) and $h = 1000$ km (right) for the same 500-year intervals as considered in Figure 7.5. It turns out that a bandwidth of less than 500 km is not large enough to capture overall patterns because for many locations on the North American continent the number of dates within such a small radius is too low. For example, a bandwidth of 300 km can cause rapid changes of estimated intensities within only few kilometers, see the ice-free corridor in Figure 7.6 (bottom left) or the signal south of the Canadian Arctic in Figure 7.6 (center and top left). On the other hand, when a bandwidth of 800 km or more is considered, the overall results remain the same but the maps are too smoothed to discern regional patterns. In most recent time intervals, e.g., almost all regions (except Florida and the Arctic) seem to have similar population numbers if $h = 1000$ km is used, see Figure 7.6 (top right). By comparison, the bandwidth of $h = 600$ km suggested by Scott's rule seems to be a good compromise. The consideration of a smaller bandwidth might be suitable if the CARD is extended substantially (particularly in Canada and the southern US) or if a regional study in an area of high sampling intensity is attempted. In contrast, a larger bandwidth could be appropriate for a global analysis. The application of a different kernel function, such as the Gaussian kernel, is not considered in this thesis since kernels with unbounded support do not correspond well to our assumption that each date influences population intensities within a fixed radius of its site.

We additionally study the effect of using two adaptive smoothing approaches, see Section 3.4.2, although such methods also contradict our assumption that all dates from the CARD indicate population within a constant radius. On the one hand, a generalized nearest neighbor estimator is applied, which works completely analogous to the KDE in (7.3) with the difference that the bandwidth varies across the considered domain. In particular, for a number of $n \in \mathbb{N}$ dates (with corresponding sites), the bandwidth for estimating $\lambda(t)$ at some location $t \in W$ is chosen to be the distance of t to the kth nearest site, where it is often assumed that $k = \lfloor n^p \rfloor$ for some $p \in (0, 1)$, compare to (3.34). We determine location-dependent bandwidths based on the $n = 7,754$ geographically distinct sites considered in Section 7.2 and choose $p = 0.65$ as this provides results that are most similar to those obtained with a fixed bandwidth of 600 km. To allow for comparability through time, these bandwidths are then used to estimate population intensity maps for all 500-year intervals. This implies, however, that the chosen proceeding is not really adaptive since the locally varying bandwidths are not determined based on the dates of the respective interval. The second approach, a variable kernel estimator, works analogously, except that a different bandwidth is assigned to each site from the data instead of to each location of the considered domain,

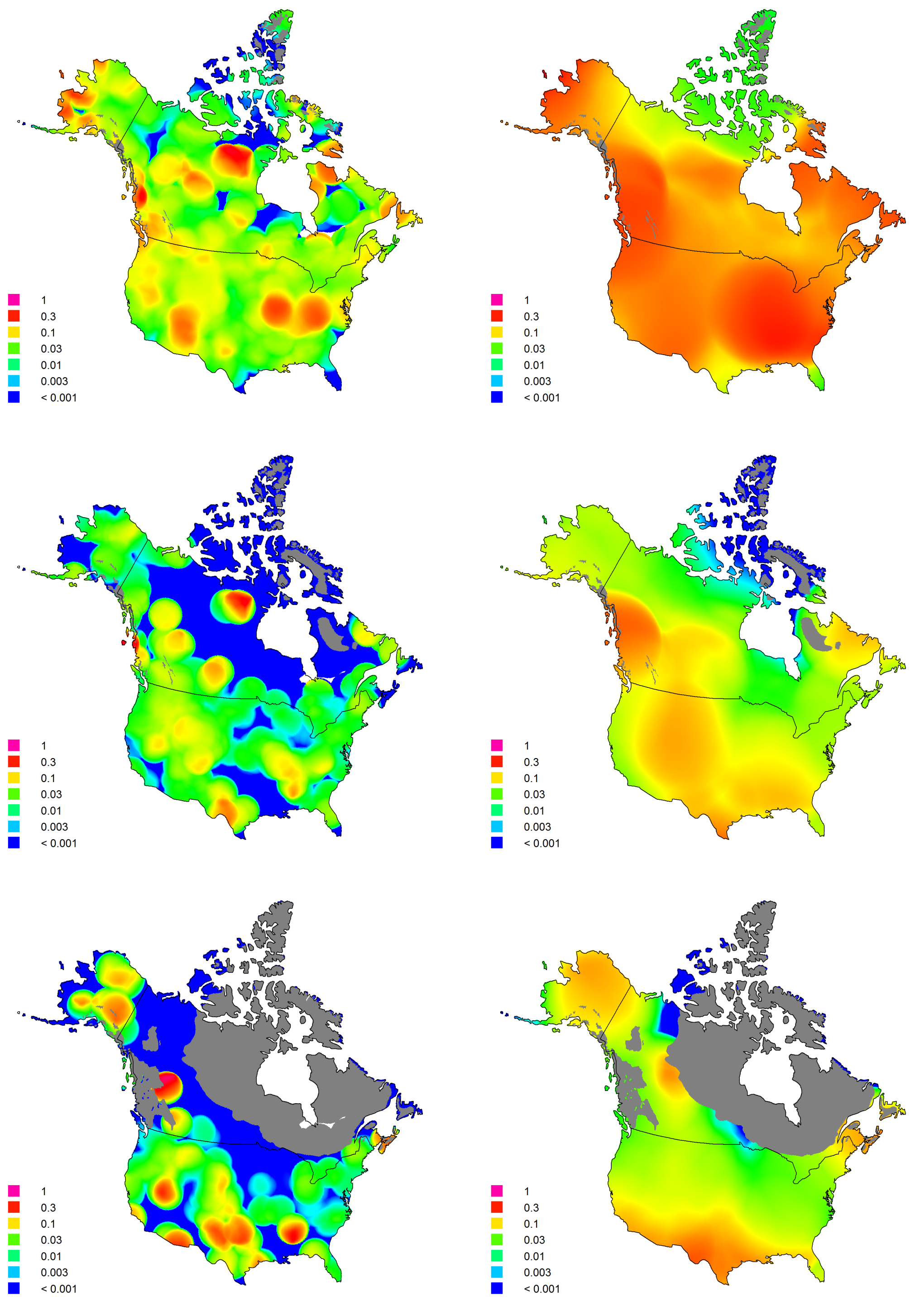


Figure 7.6.: Population intensity maps estimated using a modified KDE with a fixed bandwidth of $h = 300$ km (left) and $h = 1000$ km (right) for three selected 500-year intervals: 1600-2100 BP (top), 6400-6900 BP (center), and 12100-12600 BP (bottom). Gray colors indicate areas permanently covered by ice.

see (3.35). Similarly, for each date the corresponding bandwidth is chosen as the distance to the kth nearest site, where $k = \lfloor n^p \rfloor$ and, in our example, $n = 7,754$ and $p = 0.65$. Figure 7.7 illustrates population intensity maps for three example time intervals that were determined using the generalized nearest neighbor estimator (left) and the variable kernel estimator (right). The obtained maps show similar population estimates for both approaches. Largely smoothed patterns are revealed in regions with low site densities (e.g., central Canada) but undersmoothing occurs in more frequently sampled regions (e.g., the southeast or Wyoming). In general, however, both approaches lead to population intensity maps that show similar signals, although the spatial scale and size of single events can vary slightly. Furthermore, both sequences of maps show the same patterns as those obtained by using a globally fixed bandwidth of 600 km, which confirms the results described in Section 7.3.

7.4.2. Modification of database

To conclude this chapter, we analyze the sensitivity and robustness of estimated population intensity maps to possible errors in the underlying database. Although carefully checked, dates from the CARD are subject to well-known sources of error, e.g., dates could be contaminated with younger organic matter leading to underestimated radiocarbon ages. In particular, a total of 1,419 dates in the CARD (around 4 % of all dates) are marked as 'anomalous' (i.e., too young or old considering the archaeological context) by the contributing researchers, indicating potentially erroneous data. Such data were not removed prior to our study to avoid selectively reducing the database according to preconceived notions of paleodemographic trends. However, to determine to what extent the 'anomalous' dates in the CARD influence the resulting population patterns, we reestimate a new series of maps using the kernel approach described in Section 7.2 with a fixed bandwidth of 600 km, where all 'anomalous' dates are excluded. We find that the new maps are almost identical to those estimated based on the complete database. All population patterns remain the same throughout the entire study period at continental to regional scales. Example maps for three selected 500-year intervals are illustrated in Figure 7.8 (left), showing a high correspondence to Figure 7.5 (left).

While the analysis performed above addresses a worst-case scenario in which population intensity maps are estimated excluding dates marked as 'anomalous', we now reduce the size of the database by a much larger factor to test whether or not the CARD is spatially representative and has a sufficiently large size for all considered time intervals. For that purpose, we perform a Bernoulli experiment with a success probability of 0.5 for each date (independently for different dates) and remove all dates, where the experiment shows success. This results in a total of 16,894 dates from 5,787 geographically distinct sites. New population intensity maps are then estimated based on the reduced dataset

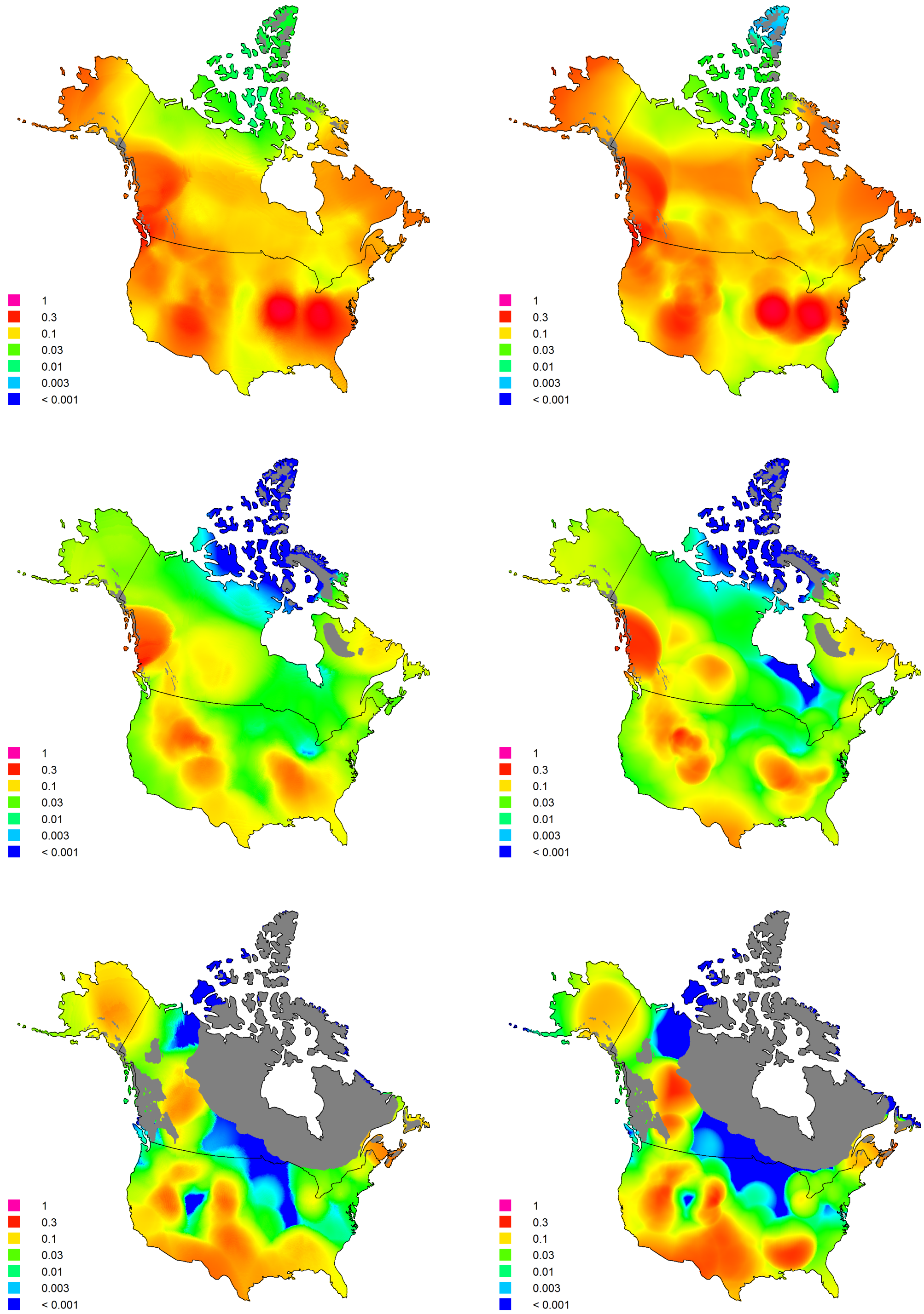


Figure 7.7.: Population intensity maps estimated using a modified generalized nearest neighbor estimator (left) and a modified variable kernel estimator (right) for three selected 500-year intervals: 1600-2100 BP (top), 6400-6900 BP (center), and 12100-12600 BP (bottom). Gray colors indicate areas permanently covered by ice.

using the kernel estimator proposed in Section 7.2 with a bandwidth of 600 km. A comparison with the original maps discussed in Section 7.3 reveals that, despite the smaller number of dates, the spatial population patterns remain mostly the same, even during earlier intervals, where the number of dates is relatively small. Only a few regional differences are visible, see, e.g., the three example maps illustrated in Figure 7.8 (right). Between 12600 BP and 12100 BP, we now observe slightly higher population intensities at the Californian coast and in Florida, whereas between 6900 BP and 6400 BP, the reduced dataset leads to slightly lower intensities in the eastern US and western Canada, compare Figure 7.5 (center and bottom left) with Figure 7.8 (center and bottom right). In more recent time periods, where more data are available, maps from the complete and the reduced dataset are almost identical, see, e.g., Figure 7.5 (top left) and Figure 7.8 (top right).

Based on the two performed robustness analyses, we conclude that the CARD is sufficiently complete for continental-scale spatio-temporal paleodemographic analyses and that individual dates as well as non-systematic errors (e.g., ‘anomalous’ dates) do not have an impact on obtained results of paleodemographic patterns.

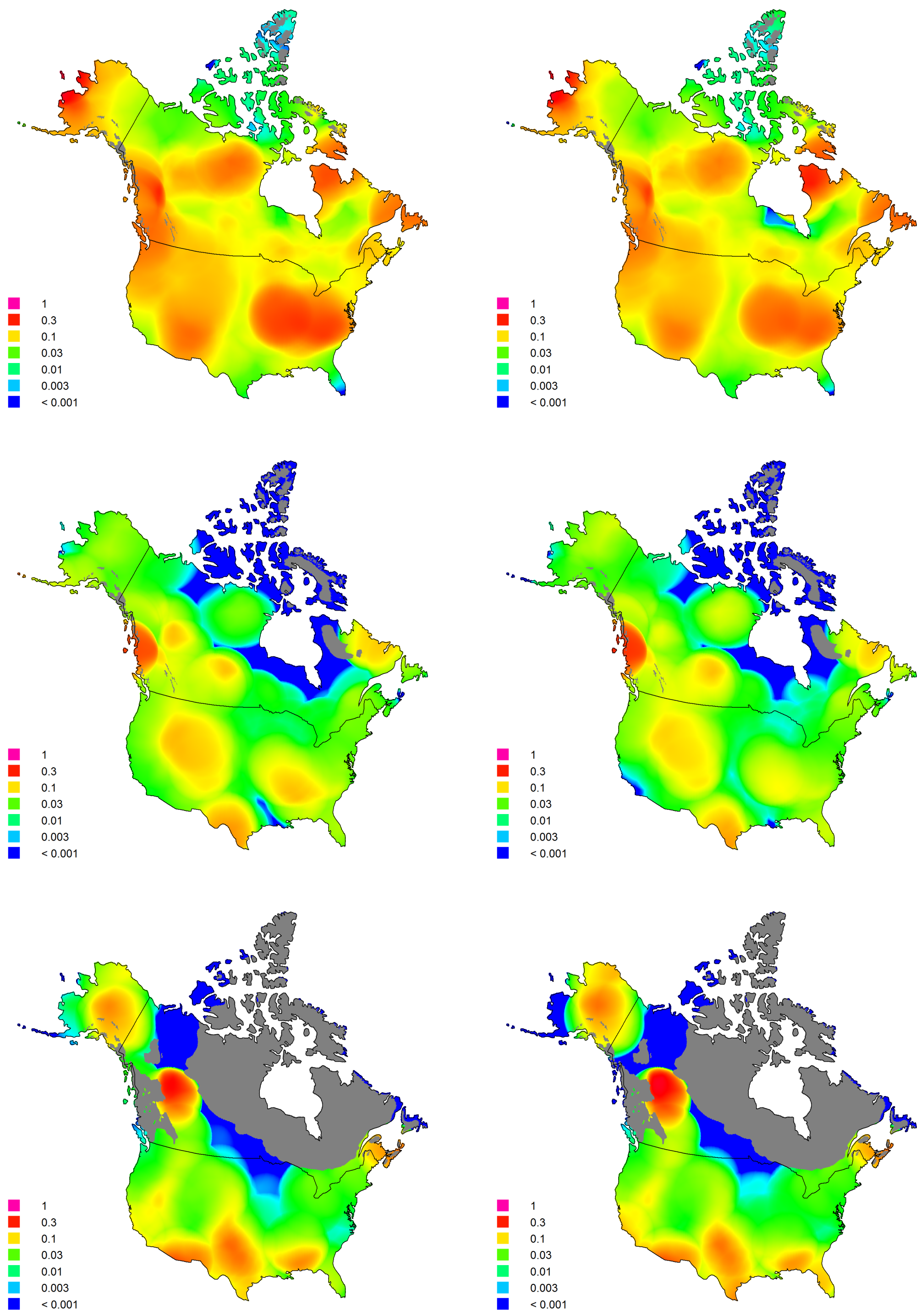


Figure 7.8.: Population intensity maps estimated using a modified KDE with a fixed bandwidth of $h = 600$ km excluding 'anomalous' dates (left) and using 50 % of dates only (right) for three selected 500-year intervals: 1600-2100 BP (top), 6400-6900 BP (center), and 12100-12600 BP (bottom). Gray colors indicate areas permanently covered by ice.

8. Spatio-temporal distribution of prehistoric vegetation abundances in North America

In order to be able to estimate spatio-temporal correlations of human populations and vegetation composition in North America during the past 13000 years, it remains to derive maps of estimated vegetation intensities. While the statistical methodology developed in Chapter 7 is, to our knowledge, the first attempt to the estimation of population intensities on a continental scale, spatial vegetation maps for (parts of) North America have been produced several times in the literature, see, e.g., Williams et al. (2004) or Paciorek and McLachlan (2009). However, on the one hand, these maps are computed based on different statistical approaches than those used in Chapter 7, making a comparison to estimated population intensity maps difficult. On the other hand, paleoecological databases have rapidly increased in size and quality during the last decade, motivating the estimation of updated vegetation intensity maps. In Section 8.1, we present the extensive database used to obtain estimates of vegetation abundances and discuss a simplified method to the calibration of radiocarbon ages. Section 8.2 then introduces an intuitive and simple approach to temporal smoothing and interpolation of pollen abundances to the target ages needed for the estimation of spatially smoothed vegetation intensity maps in Section 8.3. Section 8.4 concludes the chapter with an illustration of example maps for one taxon (*Quercus*) and a brief comparison to existing literature. The results presented in this chapter have been incorporated in Kriesche et al. (2017a).

8.1. Description of pollen data

For the purpose of reconstructing past vegetations, it is generally acknowledged to analyze prehistoric pollen records, which are considered to be an indicator of past vegetation abundance, see Birks and Birks (1980). Fossil pollen samples are typically acquired through cores taken at the bottoms of lakes and ponds, where pollen are effectively preserved in the sediment. The resulting information are collected in comprehensive paleoecological and paleoenvironmental databases.

8.1.1. The Neotoma Paleoecology Database

Spatio-temporal data of prehistoric pollen abundance used in this study are obtained from the Neotoma Paleoecology Database, see Grimm (2008). Neotoma is a comprehensive and complex compilation of fossil data from the Holocene, Pleistocene, and Pliocene (covering the last 5.3 million years) from more than 8,400 sites worldwide. It was merged from the Global Pollen Database, FAUNMAP (containing fossil mammal data), the North American Plant Macrofossil Database, and a fossil beetle database by Eric C. Grimm from the Illinois State Museum. Neotoma is publicly available and updated regularly containing data that were obtained from various sources such as pollen, plant macrofossils, vertebrate fauna, insects or mollusks. Thus, the database is used by researchers from several fields such as paleoecology, paleoclimatology, paleontology, paleoenvironmental sciences, archeology or evolutionary biology. Since many different researchers contribute to Neotoma individually, even data of the same type (pollen, mammals, etc.) are not always comparable due to different ways of acquiring and processing data. In addition, entries in Neotoma are sometimes incomplete, missing important attributes or descriptions.

8.1.2. Calibration of radiocarbon ages

For the estimation of vegetation intensities the ages of fossil pollen samples are required, which are typically determined as follows. At each site (a lake or a pond) on the North American continent with available pollen data in Neotoma, a sediment core was taken by the contributing researcher and samples were collected for a sequence $d_0 < \ldots < d_n$ of $n+1$ depths with $n \in \mathbb{N}$. Usually, each sample contains a certain number of preserved pollen, which are identified, classified, and counted. Next, the age of the sample at each depth needs to be determined. For that purpose, the technique of radiocarbon dating again is the most popular choice, see Section 7.1. However, in general a large number of samples is taken at each site and since radiocarbon dating is rather expensive, it is not used to estimate the ages of all samples. Instead, radiocarbon ages are only determined for a subset $\{d_{i_1}, \ldots, d_{i_j}\} \subset \{d_0, \ldots, d_n\}$ for $j \in \{1, \ldots, n\}$ and ages for depths $d \in \{d_0, \ldots, d_n\} \setminus \{d_{i_1}, \ldots, d_{i_j}\}$ are computed using model-based interpolation methods, see the rich literature on stochastic age-depth modeling, e.g., Bronk Ramsey (2008), Haslett and Parnell (2008) or Blaauw and Christen (2011). Next, ages obtained from the radiocarbon method (or interpolation) need to be calibrated to correct systematic errors from radiocarbon dating. Unfortunately, this is only done by a minority of researchers contributing data to Neotoma. Thus, a huge amount of radiocarbon ages needs to be calibrated, which cannot be done manually using standard calibration curves such as IntCal13, see Reimer et al. (2013). As an alternative, we suggest to convert radiocarbon ages into calibrated ages based on a smoothed radiocarbon calibration curve introduced in Grimm (2008), Figure 3. This simplified approach is not exact but in Grimm (2008) it is estimated that with a probability of 0.47 the occurring error is less than 25 years, with a probability of 0.86 the error is less than 100 years, and with a probability of 0.97 the occurring error is less than 200 years. Since this is clearly below the temporal scale of our study, it is extremely unlikely that a significant bias is introduced by using the simplified calibration curve. The result is a sequence $a_0, \ldots, a_n$ of calibrated radiocarbon ages (the unit being calibrated years BP; we simply write BP to provide consistency to Chapter 7) that correspond to the samples taken at depths $d_0, \ldots, d_n$.

The absence of calibrated ages for most samples in the Neotoma Database causes another potential problem that needs to be addressed. The correct procedure when contributing data to Neotoma would be to calibrate radiocarbon ages at depths $d_{i_1}, \ldots, d_{i_j}$ first, resulting in calibrated ages $a_{i_1}, \ldots, a_{i_j}$, and to determine ages for depths $d \in \{d_0, \ldots, d_n\} \setminus \{d_{i_1}, \ldots, d_{i_j}\}$ afterwards based on $a_{i_1}, \ldots, a_{i_j}$ using model-based interpolation methods. For the majority of sites, however, ages for depths $d \in \{d_0, \ldots, d_n\} \setminus \{d_{i_1}, \ldots, d_{i_j}\}$ are first interpolated based on uncalibrated ages at $d_{i_1}, \ldots, d_{i_j}$ before contributing data into Neotoma, and we then calibrate the ages of all $n+1$ depths $d_0, \ldots, d_n$ using the smoothed calibration curve of Grimm (2008). In general, this exchange of calibration and interpolation could lead to a bias in calibrated ages. Unfortunately, for those sites in Neotoma at which ages are only interpolated (without calibration), it does not seem possible to identify, which (uncalibrated) ages were obtained from radiocarbon dating and which from interpolation methods, making it impossible to eliminate this bias. To investigate whether the calibration bias is expected to significantly influence the analysis performed in this chapter, we perform the following case study. We select 22 independent test samples from the literature, each consisting of a sequence $d_0, \ldots, d_n$ of depths in cm and a sequence $\tilde{a}_0, \ldots, \tilde{a}_n$ of corresponding uncalibrated radiocarbon ages, with $n \in \mathbb{N}$ varying between 4 and 17. For each test sample, we determine the sequence $\{d'_1, \ldots, d'_k\} \subset [d_0, d_n]$ with $k \in \mathbb{N}$, which contains all depths between d_0 and d_n being a multiple of 5 cm. We first compute radiocarbon ages for depths $d'_1, \ldots, d'_k$ by applying linear interpolation based on $\tilde{a}_0, \ldots, \tilde{a}_n$ and afterwards calibrate interpolated ages using the smoothed calibration curve. Next, we first calibrate $\tilde{a}_0, \ldots, \tilde{a}_n$ and then determine calibrated ages for $d'_1, \ldots, d'_k$ by applying linear interpolation. This results in two calibrated radiocarbon ages a and a' for each depth $d' \in \{d'_1, \ldots, d'_k\}$, i.e., we obtain two depth-age-curves for each of the 22 test samples that can be compared to analyze the effect of exchanging calibration and interpolation. The depth-age-curves for three test samples are illustrated in Figure 8.1, showing that calibrated ages obtained with the different approaches have similar values. To provide a better overview of all 22 test samples, Figure 8.2 depicts a histogram of the differences $a - \tilde{a}$ for all depths and test samples (altogether about 1,500 values). We observe that the mean error is at -10 years, that the large majority of errors are between -100 and 100 years, and that errors larger than 300 years (or smaller than -300 years) occur extremely rarely. Moreover, most

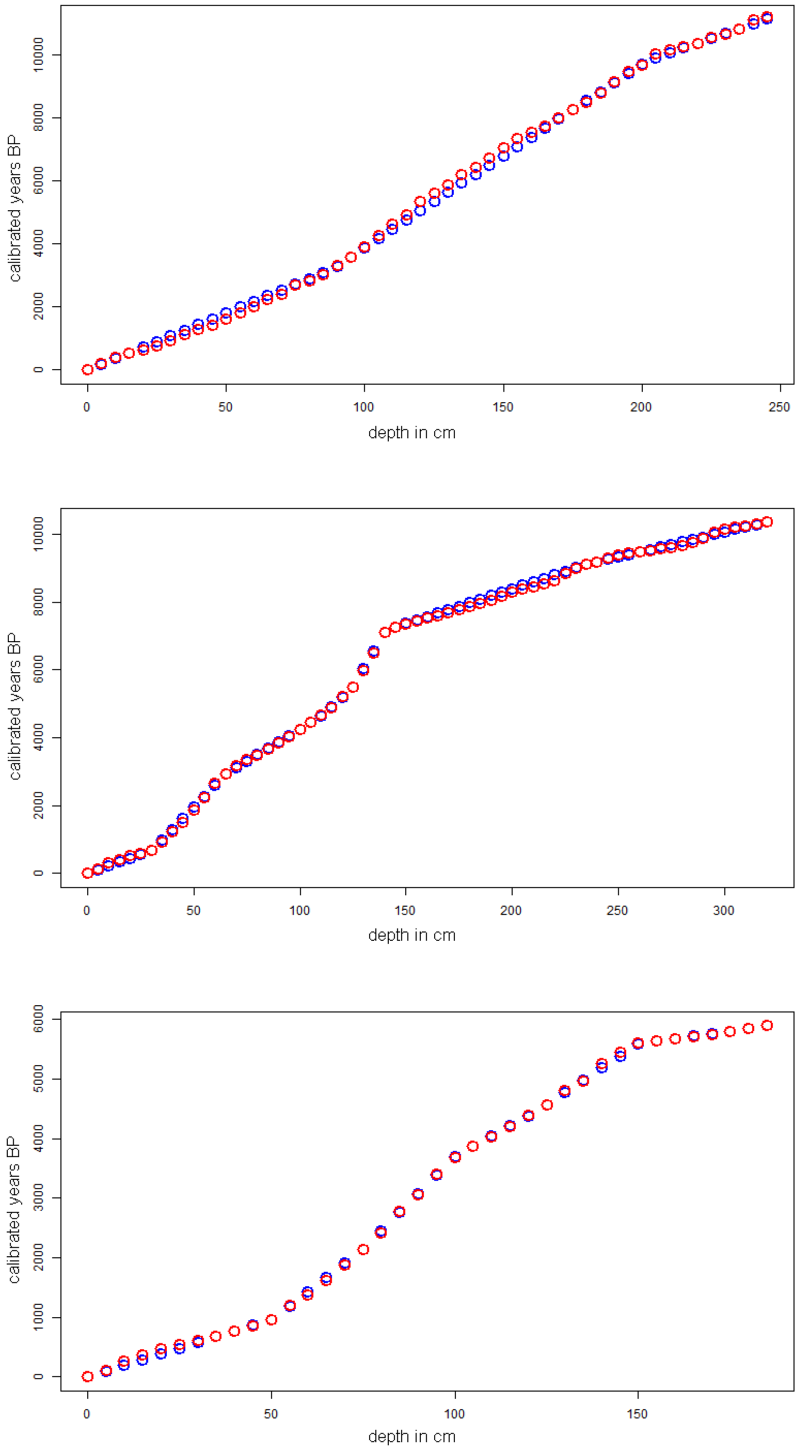


Figure 8.1.: Depth-age-curves illustrating the differences between correctly computing calibrated radiocarbon ages (first calibration, then interpolation; blue) and exchanging calibration and interpolation (red) for three selected test samples.

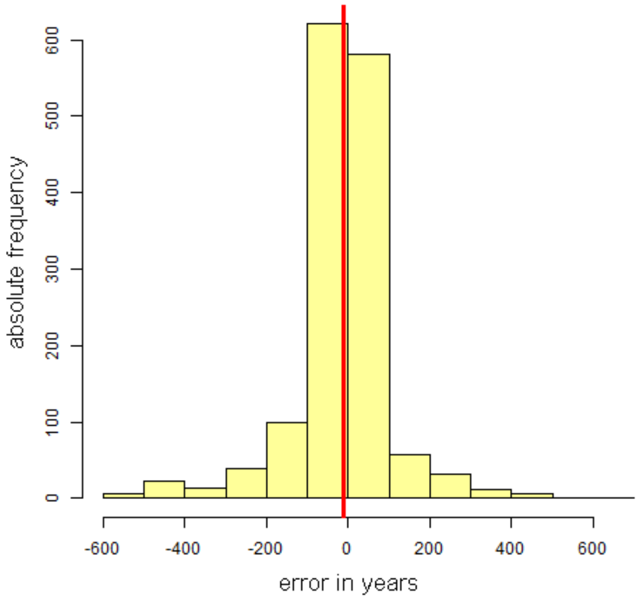


Figure 8.2.: Histogram of errors occurring when exchanging calibration and interpolation of radiocarbon ages for all chosen depths in a sequence of 22 test samples. The red line indicates the mean error.

of the larger errors occur for ages older than 13000 BP, which are not considered in our analysis. We conclude that, since the observed differences are small compared to the temporal scale of this study, errors occurring from exchanging calibration and interpolation of radiocarbon ages can be considered as negligible in the following.

8.1.3. Data selection and processing

In order to access the Neotoma Database for automatic data selection and processing, we use the R package `neotoma`, see Goring et al. (2015). The package `neotoma` uses an application programming interface to send data requests to the Neotoma Database and then forms data objects that are compatible with existing R packages for reconstruction, manipulation, presentation, processing, and inference of paleoenvironmental and paleoecological data. Three levels of information are accessible via `neotoma`, which are denoted as `sites`, `datasets`, and `downloads`. An overview of these levels and their interrelationships is given in Goring et al. (2015), Figure 1. A `site` is the most basic form of spatial information representing geographical coordinates, name, and description of excavation sites in Neotoma. According to the scope of our study, we select all `sites` which are labeled with the geopolitical id 'Canada' or 'United States', all other `sites` are dropped. Neotoma's `datasets` are collections of samples of the same type from a fixed site. Thus, in the next step all `datasets` associated with the `sites` selected before are loaded (each `dataset` belongs to a `site` but a `site` can have more than one `dataset`, e.g., one of type 'pollen' and one of type 'vertebrate fauna'). The most important attributes of a `dataset` are the collection type (describing the way data are collected) and the dataset type. Other attributes such as principal investigators or submission dates are not of interest in our study. We select all `datasets` whose collection type is equal to 'core' or 'composite' (combination of several adjacent cores) and whose dataset type is equal to 'pollen'. The most complex level of data is given by the `downloads`, which contain the most important information such as ages of samples and pollen counts. In particular, each `download` that is associated with a pollen dataset contains a sequence $d_0, \dots, d_n$ of depths and the corresponding sequence of (in most cases uncalibrated) radiocarbon ages together with the pollen counts for different taxa. If ages are not calibrated, this is done using the smoothed calibration curve of Grimm (2008), which results in ages $a_0, \dots, a_n$, see Section 8.1.2. We load all `downloads` that correspond to the `datasets` selected above and delete those `downloads` with

1. less than two depths/ages available (i.e., $n < 1$),
2. the age of at least one depth missing,
3. at least one age without characterization of age type (e.g., calibrated or not),
4. ages having different age types.

Finally, it remains to identify which taxa of a `download` should be taken into account for application in this study. A taxon and the corresponding count data are characterized by different attributes such as the taxon name, the taxon group or the variable element and unit describing the considered element, part or organ of the taxa and the unit in which the count data are measured. For each `download`, we select those taxa whose taxon group is 'vascular plants', whose variable element is equal to 'pollen' or 'spore' and whose variable unit is 'NISP' (number of identified specimen). Since data are contributed to Neotoma by many different researchers and since most taxa can be subdivided into various sub-taxa, a standardization of count data across sites is necessary. For example, one analyst might discriminate sub-genera of *Picea* (spruce), such as *Picea glauca* (white spruce) or *Picea mariana* (black spruce), while another might simply identify *Picea* to the genus level. To provide comparability, the standardization list suggested in Williams and Shuman (2008) is used, which aggregates count data for up to 64 taxa. Finally, we observe that comparing pollen counts (for a fixed taxon) across time and space is still problematic due to, e.g., spatially and temporally varying sampling effort. Thus, the usual practice is to compute relative pollen abundances for all sites, ages, and taxa, which are much easier to compare and interpret. Altogether, we obtain

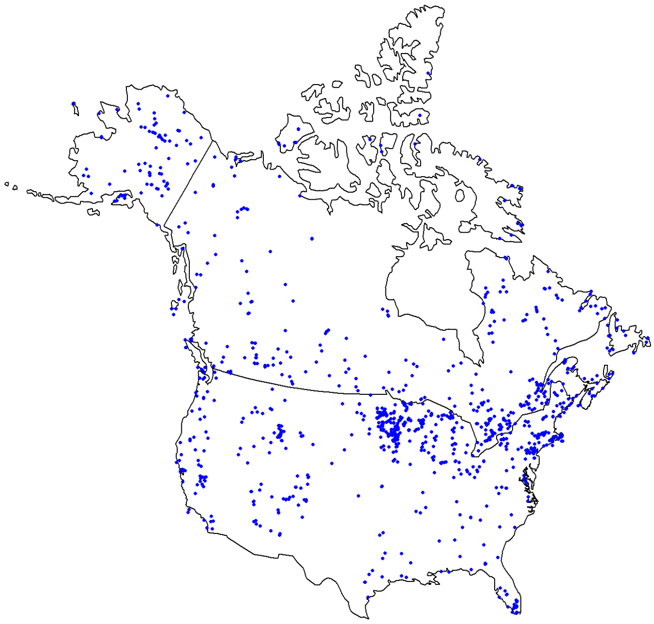


Figure 8.3.: Sites with data of relative pollen abundances in the Neotoma Paleoecology Database.

a set of $1,151$ sites, which are distributed inhomogeneously over the entire North American continent, see Figure 8.3. In particular, we observe a large number of sites in southeastern Canada, the northeastern and western US and Alaska, whereas northern Canada and the southern US show low site densities. Each of the depicted sites contains a sequence $a_0, \dots, a_n$ of calibrated radiocarbon ages together with the corresponding relative pollen abundances of 64 taxa. In this thesis, pollen data were compiled for 10 selected taxa, which are major constituents of the forests and prairies of North America and some of which were extensively used by First Nations people. Those are *Acer* (maple), *Carya* (hickory), *Castanea* (chestnut), *Fagus* (beech), *Juglans* (walnut and butternut), *Picea* (spruce), *Pinus* (pine), Poaceae (grasses), *Populus* (includes poplar, aspen and cottonwood), and *Quercus* (oak).

8.2. Temporal interpolation and smoothing of pollen abundances

To allow for the estimation of spatial vegetation intensity maps, relative pollen abundances need to be available at (a subset of) the $1,151$ sites shown in Figure 8.3 simultaneously for the same years. However, the ages $a_0, \dots, a_n$ for which pollen data are obtained from Neotoma are different from site to site, which makes the application of interpolation or smoothing methods necessary. Since relative pollen abundances may vary considerably even during short time periods, which is most likely a result of random sampling errors, it is recommended to use smoothing methods over interpolation if possible, see Takezawa (2006). In the following, we consider a fixed site from the Neotoma database with available pollen data and a fixed taxon (one of the 10 described in Section 8.1.3). With $(a_0, p_0), \ldots, (a_n, p_n)$, where $0 \leq a_0 < \ldots < a_n$ and $p_i \in [0,1]$ for all $i = 0, \ldots, n$, we denote the ages of the available samples in years BP (i.e., a_n describes the age of the oldest sample) and the corresponding relative pollen abundances of the chosen taxon. A random design approach seems suitable in this context, see Section 3.4.3, since the ages of samples are not fixed a priori by the analyst collecting the data and can thus be considered to be random. Furthermore, samples for different ages are acquired and investigated independently. Accordingly, we interpret $(a_0, p_0), \ldots, (a_n, p_n)$ as (sorted) realizations of some independent and identically distributed absolutely continuous random vectors $(A_0, P_0), \ldots, (A_n, P_n)$ taking values in $[0, \infty) \times [0, 1]$. Furthermore, we consider the PDF $f_A : \mathbb{R} \to [0, \infty)$ of A_0. It seems almost impossible to find a parametric representation describing the relationship between the random pollen abundances $P_0, \ldots, P_n$ and the random ages $A_0, \ldots, A_n$ sufficiently well, which is why we generally assume that

$$P_i = p(A_i) + v^{\frac{1}{2}}(A_i)\,\varepsilon_i, \quad i = 0, \ldots, n,$$

where $p : \mathrm{supp}(f_A) \to [0, 1]$ with

$$p(a) = \mathbb{E}\,(P_0 \mid A_0 = a), \quad a \in \mathrm{supp}(f_A),$$

and $v : \mathrm{supp}(f_A) \to [0, \infty)$ with

$$v(a) = \mathrm{var}\,(P_0 \mid A_0 = a), \quad a \in \mathrm{supp}(f_A),$$

are the conditional expectation function and the conditional variance function of P_0 given A_0 and $\varepsilon_0, \ldots, \varepsilon_n$ denote some random variables, called residuals, with $\mathbb{E}\,\varepsilon_i = 0$ and $\mathrm{var}\,\varepsilon_i = 1$ for all $i = 0, \ldots, n$.

The conditional expectation function p can be estimated using kernel smoothing. For that purpose, we consider the time intervals $I_1, \ldots, I_n$, where $I_i = [a_{i-1}, a_i)$ for $i = 1, \ldots, n-1$ and $I_n = [a_{n-1}, a_n]$, and by $h_i = a_i - a_{i-1}$ we denote the length of interval I_i for $i = 1, \ldots, n$. In order to allow for a smooth estimation of p, the bandwidth h, which controls the degree of smoothing in one-dimensional kernel estimators, should be chosen not smaller than the maximum $\max\{h_1, \ldots, h_n\}$. However, for some sites and taxa, a long period I_i without a sample occurs, which leads to a bandwidth that can cause oversmoothing in other intervals by eliminating too many details. To avoid such effects, we only consider those intervals $\{I_{i_1}, \ldots I_{i_k}\} \subset \{I_1, \ldots, I_n\}$ with $k \in \{1, \ldots, n\}$ that have a length of not more than 2000 years. Furthermore, we define

$I = I_{i_1} \cup \ldots \cup I_{i_k}$ and $h = \max\{h_{i_1}, \ldots, h_{i_k}\}$, where it is assumed that $I \subset \operatorname{supp}(f_A)$. Then, an estimate $\hat{p}$ of p can be determined using a (one-dimensional) NWE with bandwidth h according to (3.41), i.e.,

$$\hat{p}(a) = \frac{\sum_{i=0}^{n} p_i \, \kappa^G \left(\dfrac{a - a_i}{h} \right)}{\sum_{i=0}^{n} \kappa^G \left(\dfrac{a - a_i}{h} \right)}, \quad a \in I, \tag{8.1}$$

with $\kappa^G : \mathbb{R} \to [0, \infty)$ being the one-dimensional Gaussian kernel defined in (3.31). Motivation for choosing the Gaussian kernel over, e.g., the Epanechnikov kernel is that it has an unbounded support and is thus expected to provide smooth estimates even in regions of few data points. Furthermore, we avoid that the denominator in (8.1) can be equal to zero. However, for $a \notin I$, using the NWE with bandwidth h as chosen above occasionally results in sudden decreases or increases making it an inappropriate choice. Therefore, we alternatively suggest to estimate p using linear interpolation, which leads to an estimate $\tilde{p}$ of p defined by

$$\tilde{p}(a) = \sum_{i=1}^{n} \left(p_{i-1} + (a - a_{i-1}) \frac{p_i - p_{i-1}}{a_i - a_{i-1}} \right) \mathbb{1}_{I_i}(a), \quad a \in [a_0, a_n]. \tag{8.2}$$

Finally, we set $\hat{p}(a) = \tilde{p}(a)$ for $a \in [a_0, a_n] \setminus I$.

The proposed methodology is applied to 10 selected taxa, see Section 8.1.3. Figure 8.4 illustrates data of relative pollen abundances from Neotoma for three sites together with estimates $\{\hat{p}(a), a \in I\}$ and $\{\tilde{p}(a), a \in [a_0, a_n]\}$ obtained using the NWE in (8.1) and linear interpolation according to (8.2), respectively. We find that choosing the smoothing parameter h as explained above leads to estimates that capture the temporal development very precisely, including abrupt in- or decreases, see, e.g., the yellow curve in Figure 8.4 (top). On the other hand, noise is also eliminated quite well, see, e.g., the green curve in Figure 8.4 (center) and the blue curve in Figure 8.4 (bottom). In Figure 8.4 (bottom), we furthermore have that $I \neq [a_0, a_n]$ since for ages between 14300 BP and 18000 BP interpolation is preferred over smoothing due to missing data. We also point out that, although we will keep the notation 'relative pollen abundance' in the following, estimates $\hat{p}(a)$ for all 64 taxa at a fixed age $a \in [a_0, a_n]$ do not longer sum to one in general.

8.3. Vegetation intensity maps

Now that estimated relative pollen abundances are available at a large number of sites in North America simultaneously for the same years, we address the question how

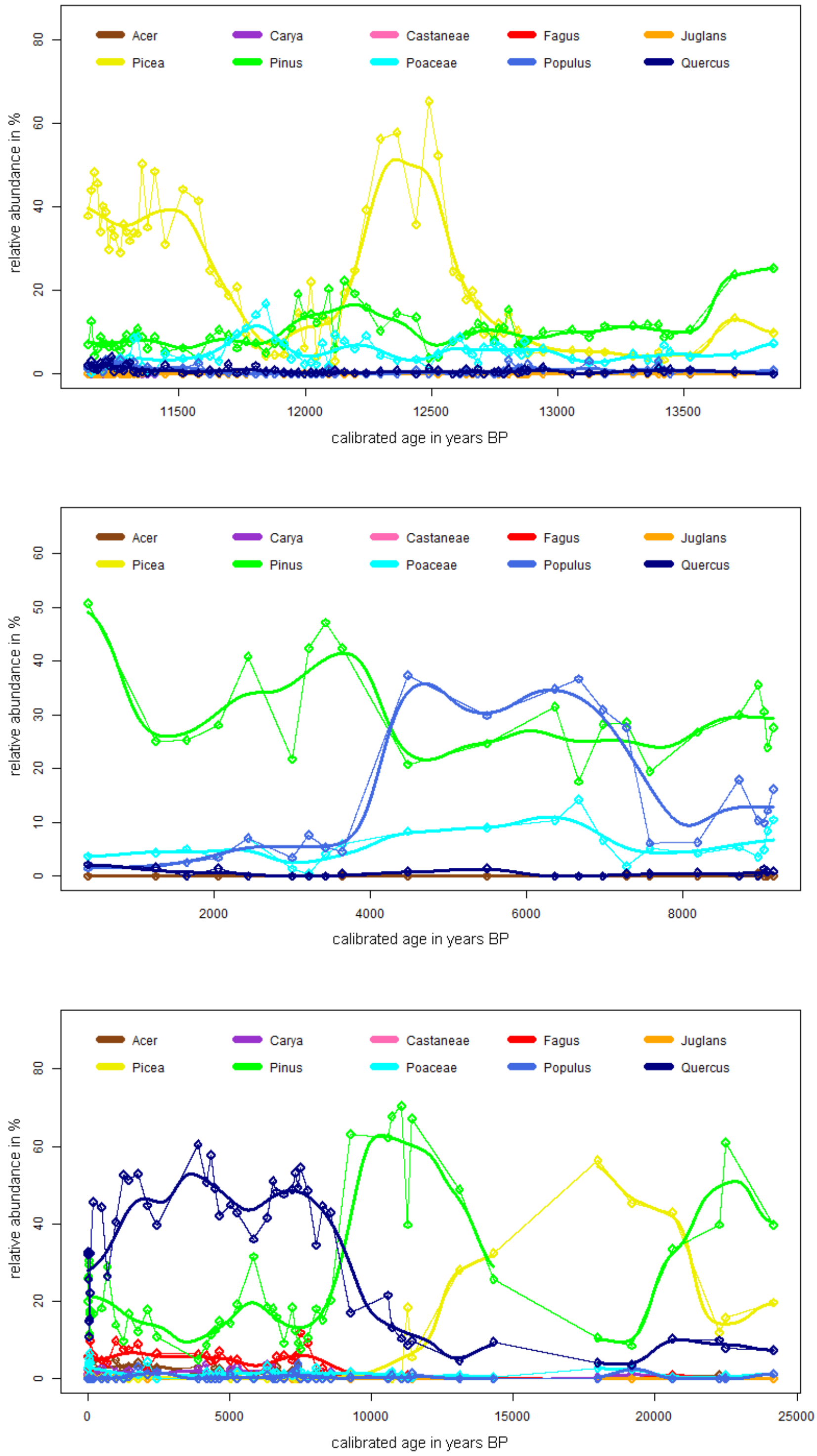


Figure 8.4.: Relative pollen abundances for three different sites and 10 selected taxa: data from Neotoma (points), estimates obtained using linear interpolation (thin lines), and estimates obtained using a NWE (bold lines).

smooth vegetation intensity maps can be determined. For that purpose, we use similar nonparametric kernel methods as applied in Section 7.2 for the estimation of population intensity maps. Furthermore, we describe an approach to estimate the (temporally changing) ranges of the considered taxa based on obtained vegetation intensities.

8.3.1. Nonparametric estimation of vegetation intensity maps

In order to compute smooth vegetation intensity maps that are comparable to population intensity maps estimated in Chapter 7, we consider the sequence of years 750 BP, 850 BP, $\ldots$, 12750 BP, which correspond to the midpoints of the 500-year intervals $500 - 1000$ BP, $600 - 1100$ BP, $\ldots$, $12500 - 13000$ BP introduced in Section 7.2. The estimation procedure proposed in the following is applied to each of the 121 years and each taxon specified in Section 8.1.3 separately. Let y be a fixed year, i.e., $y \in \{750, 850, \ldots, 12750\}$, and let $W \subset \mathbb{R}^2$ again denote the compact set of all locations on the North American continent that are not permanently covered by ice in year y (according to reconstructions of prehistoric ice sheets, see Dyke, 2002). We only consider those of the 1,151 sites shown in Figure 8.3 with $y \in [a_0, a_n]$, where a_0 and a_n are the site-specific minimal and maximal available calibrated radiocarbon ages introduced in Section 8.1.2. At all other sites no pollen abundances are available for year y. This results in a sequence $(s_1, \pi_1), \ldots, (s_m, \pi_m)$, with $m \leq 1,151$, where $s_1, \ldots, s_m$ denote all sites in North America with smoothed or interpolated pollen abundances $\pi_1, \ldots, \pi_m$ in year y that were derived according to (8.1) or (8.2) (for time intervals with sparse data).

As in some time periods relative pollen abundances vary considerably (even among closely located sites) due to statistical noise and as we are interested in large scale patterns of spatial vegetation intensities, it seems more suitable to apply nonparametric spatial smoothing methods than interpolation methods such as kriging, see, e.g., Wackernagel (2003), Diggle and Ribeiro Jr. (2007) or Cressie and Wikle (2011). Furthermore, we aim to provide as many similarities to the estimators applied in Section 7.2 for the computation of population intensity maps as possible. We again choose a random design approach, see Section 3.4.3, as sampling sites were not designed by one analyst but chosen individually by many different researchers and can thus be considered to be independent random vectors. Furthermore, corresponding pollen abundances are estimated independently for different sites based on the data from Neotoma, which are also sampled independently across sites. Accordingly, we suppose that $(s_1, \pi_1), \ldots, (s_m, \pi_m)$ can be interpreted as realizations of some independent and identically distributed absolutely continuous random vectors $(S_1, \Pi_1), \ldots, (S_m, \Pi_m)$ with values in $W \times [0, 1]$. By $f_S : W \to [0, \infty)$ we denote the PDF of S_1, where we assume that $f_S(t) > 0$ for all $t \in W$. Again, it seems impossible to find a parametric representation that models the relationship between sites $S_1, \ldots, S_m$ and relative pollen abundances $\Pi_1, \ldots, \Pi_m$ sufficiently well, which is why we suppose that

$$\Pi_i = \pi(S_i) + \tilde{v}^{\frac{1}{2}}(S_i)\,\tilde{\varepsilon}_i, \quad i = 1, \ldots, m,$$

where $\pi : W \to [0, 1]$ with

$$\pi(t) = \mathbb{E}\,(\Pi_1 \mid S_1 = t), \quad t \in W,$$

and $\tilde{v} : W \to [0, \infty)$ with

$$\tilde{v}(t) = \mathrm{var}\,(\Pi_1 \mid S_1 = t), \quad t \in W,$$

denote the conditional expectation function and the conditional variance function of Π_1 given S_1 and $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_m$ are some random residuals with $\mathbb{E}\,\tilde{\varepsilon}_i = 0$ and $\mathrm{var}\,\tilde{\varepsilon}_i = 1$ for $i = 1, \ldots, m$. Since pollen are interpreted as an index of past plant abundance, see Section 8.1, we consider the field $\{\pi(t), t \in W\}$ as a map of expected vegetation intensities of the considered taxon in year y. We suggest to estimate $\{\pi(t), t \in W\}$ based on $(s_1, \pi_1), \ldots, (s_m, \pi_m)$ using a two-dimensional NWE, see Section 3.4.3. According to (3.41), an estimate $\{\hat{\pi}(t), t \in W\}$ of $\{\pi(t), t \in W\}$ can be determined by

$$\hat{\pi}(t) = \frac{\sum_{i=1}^{m} \pi_i\, \kappa_h^E(t, s_i)}{\sum_{i=1}^{m} \kappa_h^E(t, s_i)}, \quad t \in W, \tag{8.3}$$

where for $h > 0$ the function $\kappa_h^E : W \times W \to [0, \infty)$ denotes the (scaled) two-dimensional radially symmetric Epanechnikov kernel with bandwidth matrix $H = \mathrm{diag}(h^2, h^2)$ given in (7.1). Furthermore, we suggest to set $h = 600$ km to ensure comparability to the population intensity maps estimated in Section 7.2, which is also the reason for choosing the Epanechnikov kernel over a kernel with unbounded support (such as the Gaussian kernel). However, in contrast to the estimation of population intensity maps, we do not need to account for sampling biases, boundary effects, and errors occurring due to taphonomic loss. On the one hand, the denominator in (8.3) prevents the estimate from being influenced by inhomogeneous sampling strategies and boundary effects. On the other hand, relative pollen abundances do not contain taphonomic biases as it can be assumed that pollen do not degrade over the time-scale of our study. Computing $\{\hat{\pi}(t), t \in W\}$ for all years $y \in \{750, \ldots, 12750\}$ results in a sequence of estimated vegetation intensity maps for the selected taxon.

8.3.2. Estimation of taxon ranges

Most taxa considered in this thesis have a region of typical occurrence, which is denoted as taxon range (e.g., the eastern and southern US for *Quercus*). Only in the taxon range estimated vegetation intensities have significantly positive values, whereas in the remaining regions intensities are zero or very close to zero (due to few pollen grains being transported by wind or caused by data, sampling, and measurement errors). When comparing a population and a vegetation intensity map (see Chapter 9), only the range of the selected taxon should be taken into account, as in this region the taxon represents a significant part of the local vegetation in the considered year. For example, if correlations between population activity and the intensity of *Quercus* are estimated, only the eastern and southern US should be analyzed. Outside the range of *Quercus* (e.g., in the western US and in Canada), estimated vegetation intensities will be very close to zero although population activity is likely to occur in some parts of these regions, which will dilute correlation results. For that purpose, we suggest an approach to determine (temporally varying) estimates of the taxon range based on vegetation intensities. Let $\{\hat{\pi}(t), t \in W\}$ be an estimated vegetation intensity map, which is computed according to (8.3) for a selected taxon with (unknown) range $\xi \subset W$ in a given year y. Then, we suggest to compute an estimate $\hat{\xi}$ of ξ by

$$\hat{\xi} = \{t \in W : \hat{\pi}(t) \geq u \cdot \max\{\hat{\pi}(t), t \in W\}\}\,, \tag{8.4}$$

with a suitable threshold $u \in (0, 1)$. In other words, we estimate the range as the set of all locations on the North American continent at which the local vegetation intensity is at least u times the global maximum of the vegetation intensity map. In order to determine an optimal choice for the threshold u, the taxon range is estimated for the most current year ($y = 750$ BP) using thresholds $u = 0.05, 0.1, 0.15, 0.2, 0.25, 0.3$. Then, a visual comparison to the taxon's modern range is made, see Thompson et al. (1999), where we observe that for all 10 taxa considered in the previous sections the threshold $u = 0.2$ provides the best overall match. Since the estimate $\hat{\xi}$ as defined in (8.4) always depends on the maximum of the considered vegetation intensity map (and thus changes over time), we suppose that the suggested approach is also able to capture the typical taxon ranges for older years. This will be confirmed by the example maps shown in Section 8.4.

8.4. Discussion of results for Quercus

Using the statistical methodology proposed in Sections 8.2 and 8.3, we estimated sequences of vegetation intensity maps for all 10 taxa specified in Section 8.1.3. This results in a large number of maps and discussing all of them clearly goes beyond the scope of this thesis. As an example illustrating vegetation intensity maps and taxon ranges obtained according to (8.3) and (8.4), we briefly discuss results for *Quercus* (oak). *Quercus* is the characteristic species of the eastern deciduous forest and serves as a general indicator for the extent of this ecosystem. It is an important source of food for both humans and game animals, see McShea and Healy (2002), which suggests the assumption that the abundance of *Quercus* has a significant impact on human populations. Figure 8.5 shows estimated vegetation intensity maps of *Quercus* together with the corresponding estimated taxon ranges for three selected years. The general development of *Quercus* over the past 13000 years according to estimated vegetation intensities is summarized in the following. Prior to 11000 BP, highest values of *Quercus* were restricted to Florida and the states along the eastern Gulf Coast, see Figure 8.5 (bottom). The range expanded and moved to the north and west after 11000 BP. At a smaller scale, the maps capture the lower values of *Quercus* in the upper slopes of the Appalachian Mountains, whereas *Quercus* became more abundant west of this mountain chain. During this time, *Quercus* occurred in all southern and eastern states, see Figure 8.5 (center). At around 9000 BP, *Quercus* became less abundant in the south, and had the highest values in a band across the Mid-Atlantic States. Lower vegetation intensities are also shown in the Great Lakes region and Texas. This pattern remained for the next few thousand years, see Figure 8.5 (top), followed by a slight decrease in *Quercus* between 2000 BP and 1500 BP.

Maps depicting the spatio-temporal development of tree taxa across eastern North America (e.g., for *Quercus*) have been produced several times since the 1970s. During the last decades, extensive paleoecological databases were compiled and innovative statistical methods to estimate such maps were developed, which allows to infer a detailed picture of prehistoric vegetation patterns, see, e.g., Williams et al. (2004) and the references therein. The latest version discussed in the mentioned paper is based on a previous database, the North American Pollen Database with 759 sites, to produce maps every 1000 to 2000 years using a tri-cubic distance weighting to average pollen data from a 300 km $\times$ 300 km $\times$ 500 m window to a 50 km $\times$ 50 km lattice. There is a very close visual correspondence of our maps to those of Williams et al. (2004); all of the features discussed above are seen in both sets of maps. The overall migration pattern at the scale discussed in this study is captured, only showing marginal differences due to the updated database and differences in smoothing and mapping methods. This suggests that the methodology proposed in this paper is able to provide reliable indicators of vegetation abundance, which can be used for statistical comparison to spatio-temporal human population intensities.

Besides the simple nonparametric smoothing techniques applied in this chapter or in Williams et al. (2004), a more sophisticated approach to the inference of local vegetation intensities for tree taxa is discussed in Paciorek and McLachlan (2009), where Bayesian models and methods have also been used to estimate uncertainties in obtained results. However, we do not consider this to be necessary in our study as estimated vegetation intensity maps correspond particularly well with those of the existing literature, see above. Furthermore, the proposed methodology is closely related to that we used when estimating population intensity maps in Section 7.2, which is preferred to enable a qualitative comparison of vegetation and population intensities in Chapter 9.

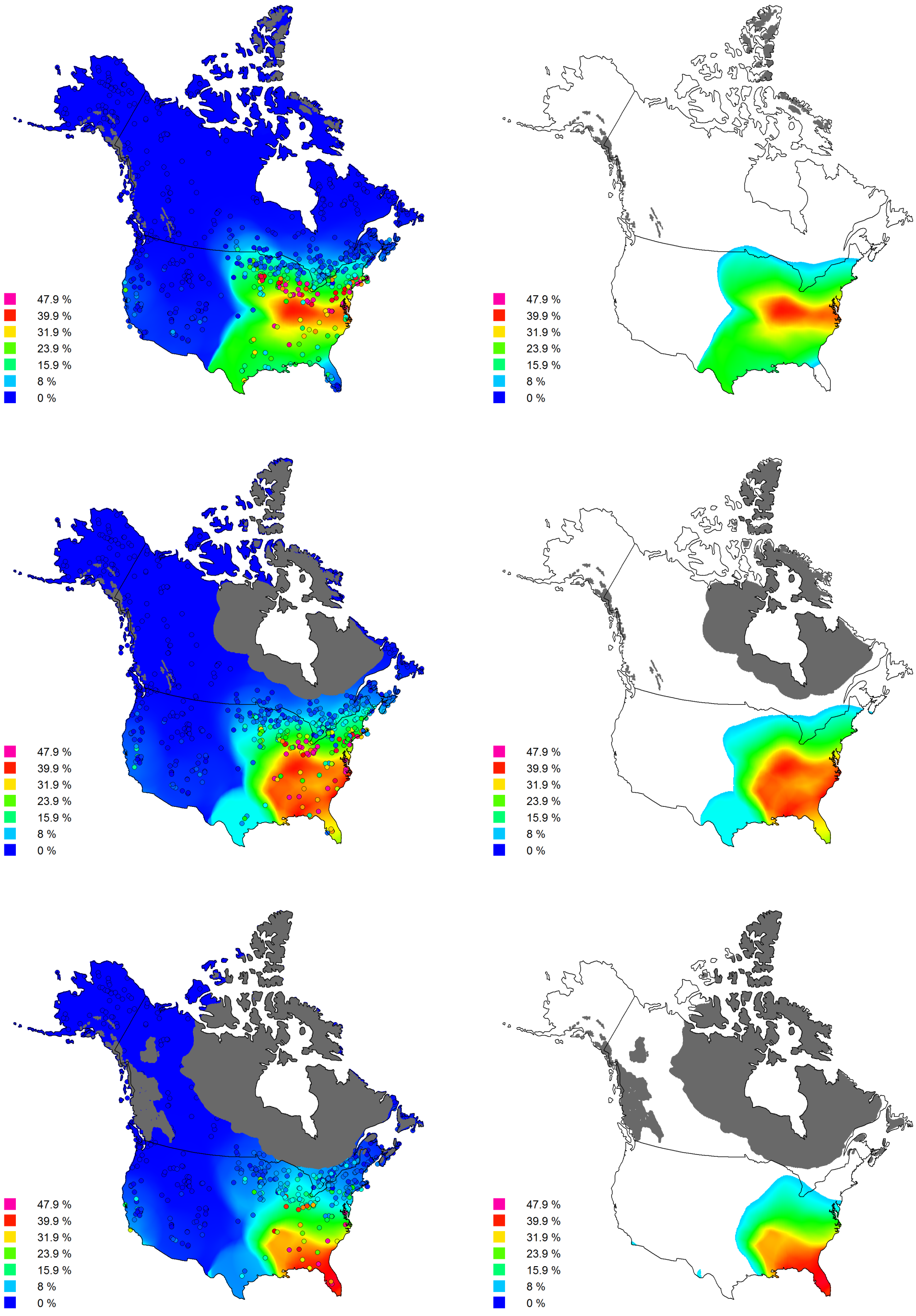


Figure 8.5.: Estimated vegetation intensity maps of *Quercus* for the years 2550 BP (top), 9450 BP (center), and 12050 BP (bottom). The left-hand side illustrates intensities for the entire continent together with temporally smoothed abundances of *Quercus* at sites from the Neotoma database. On the right, vegetation intensities are restricted to the estimated taxon range. Gray colors indicate areas permanently covered by ice.

9. Correlation analysis for vegetation abundances and population intensities

After spatio-temporal estimates of prehistoric vegetation and population intensity maps for North America have been derived based on comprehensive paleoecological and archaeological databases, we now address the analysis of interrelationships between both sets of maps. It still remains unclear if and how Native American populations interacted with their environment. On the one hand, it seems plausible that natives responded to environmental and climatic changes that have a strong impact on their living conditions. On the other hand, some researchers hypothesize that native populations directly influenced the composition of forests, e.g., by burning down trees to make the land available for agriculture and by growing plants that serve as a source of food. In order to provide a tool for the identification of such relationships, we describe a simple approach to estimate spatial cross-correlation functions of vegetation and population intensities as well as cross-correlations of changes in vegetation and population at various temporal lags, see Sections 9.1 and 9.2. Furthermore, Section 9.3 introduces a method to compute nonparametric confidence bands of estimated cross-correlation functions based on different resampling methods to assess the significance of obtained results. The chapter is concluded by a presentation and discussion of results for one taxon (*Quercus*) in Section 9.4. The contents presented in the following have been incorporated in Kriesche et al. (2017a). A more detailed discussion of correlation results for further taxa in an archaeological and paleoecological context is provided in Gajewski et al. (2017).

9.1. Cross-correlation functions of vegetation and population intensity maps

In a first step, we aim to analyze correlations between prehistoric vegetation and population intensity maps that were estimated in Chapters 7 and 8 of the present thesis. We again consider the 10 taxa introduced in Section 8.1.3 (*Acer*, *Carya*, *Castanea*, *Fagus*, *Juglans*, *Picea*, *Pinus*, Poaceae, *Populus*, and *Quercus*) and the sequence of years 750 BP, 850 BP, $\ldots, 12750$ BP, compare to Section 8.3.1. In the following, we fix one of the 10 taxa and a year $y \in \{750, 850, \ldots, 12750\}$ and by $W \subset \mathbb{R}^2$ we again denote the compact set of all locations on the North American continent that are not permanently covered by ice in year y. Next, we consider the corresponding map $\{\hat{\pi}(t), t \in \hat{\xi}\}$ of estimated vegetation intensities in year y and the map $\{\hat{\lambda}(t), t \in \hat{\xi}\}$ of estimated population intensities for the 500-year time period $[y-250, y+250]$, both being restricted to the estimated taxon range $\hat{\xi} \subset W$, compare to (7.3), (8.3), and (8.4). In estimation theory, estimators are commonly modeled as random elements, which is why $\hat{\xi}$ is considered to be a realization of some random closed subset Ξ of W, see Section 3.3.6. Furthermore, the estimated maps $\{\hat{\pi}(t), t \in \hat{\xi}\}$ and $\{\hat{\lambda}(t), t \in \hat{\xi}\}$ are interpreted as realizations of random fields $\Pi = \{\Pi(t), t \in W\}$ and $\Lambda = \{\Lambda(t), t \in W\}$ restricted to Ξ, where $\Pi(t)$ and $\Lambda(t)$ take values in $[0,1]$ and $[0,\infty)$, respectively, for each $t \in W$.

Inference of joint probabilistic properties of two random fields based on one pair of realizations is only possible if certain assumptions on the spatial dependency structure of the fields are made. We suppose in the following that Π and Λ are jointly second-order motion-invariant, see Section 3.2.4 (which means that both Π and Λ are second-order motion-invariant, too, see Section 3.2.3). This implies that Π and Λ have constant expectation functions, i.e., $\mathbb{E}\,\Pi(t) = \mu_\Pi$ and $\mathbb{E}\,\Lambda(t) = \mu_\Lambda$ for each $t \in W$, and that for any pair of locations $s, t \in W$ the covariances $\operatorname{cov}(\Pi(s), \Pi(t))$, $\operatorname{cov}(\Lambda(s), \Lambda(t))$, and $\operatorname{cov}(\Pi(s), \Lambda(t))$ only depend on the (great-circle) distance $d_{GC}(s,t)$ between s and t (due to W being a geographical domain), see Remark 3.2.4. Clearly, these assumptions are rather unrealistic on a continental scale due to, e.g., geographical and climatic differences. However, when estimating cross-correlation functions, we only consider the restriction of vegetation and population intensity maps to the estimated taxon range, which covers a smaller, geographically more homogeneous region for most taxa (in particular for *Quercus*, which is analyzed in more detail in Section 9.4).

We suggest to analyze relationships between vegetation and population intensities by estimating cross-covariance and cross-correlation functions. Let $r_0 > 0$ be defined as in (3.1), i.e., r_0 denotes the maximum distance such that for any $r \le r_0$ there is at least one pair s, t of locations in W with $d_{GC}(s,t) = r$. As Π and Λ are assumed to be jointly second-order motion-invariant, the motion-invariant cross-covariance function $C_{\Pi\Lambda} : [0, r_0] \to \mathbb{R}$ of Π and Λ is defined by

$$C_{\Pi\Lambda}(r) = \operatorname{cov}\big(\Pi(s), \Lambda(t)\big), \quad s, t \in W \text{ such that } d_{GC}(s,t) = r, \tag{9.1}$$

compare to (3.5). Cross-covariance functions are rather difficult to interpret, which is why they are typically normalized to obtain cross-correlation functions. By using that the variances $\sigma_\Pi^2 = \operatorname{var}\Pi(t)$ and $\sigma_\Lambda^2 = \operatorname{var}\Lambda(t)$ do not depend on $t \in W$, where we assume that $\sigma_\Pi^2, \sigma_\Lambda^2 > 0$, the (motion-invariant) cross-correlation function $\rho_{\Pi\Lambda}$:

Figure 9.1.: Realization of a stationary Poisson point process with intensity $\alpha = 0.0003$, the points of which are used as sample points for the estimation of cross-correlation functions.

$[0, r_0] \to [-1, 1]$ of Π and Λ is represented as

$$\rho_{\Pi\Lambda}(r) = \frac{C_{\Pi\Lambda}(r)}{\sqrt{\sigma_\Pi^2\,\sigma_\Lambda^2}}, \quad r \in [0, r_0],$$

see Remarks 3.2.3 and 3.2.6.

Even though the whole trajectories of population and vegetation intensity maps were estimated in the previous chapters, we need to choose a finite sequence $u_1, \ldots, u_k \in \hat{\xi}$ of $k \in \mathbb{N}$ sample points, at which the values $\hat{\pi}(u_1), \ldots, \hat{\pi}(u_k)$ and $\hat{\lambda}(u_1), \ldots, \hat{\lambda}(u_k)$ are determined for the estimation of cross-correlation functions. One possibility is to consider a regularly spaced lattice for this purpose but this raises the question of how to choose its origin and orientation. To avoid making an arbitrary choice here, we alternatively suggest to generate a realization of a stationary Poisson point process with intensity $\alpha = 0.0003$, see Section 3.3.3, and use those points of the process as sample points $u_1, \ldots, u_k$ that fall into the estimated taxon range $\hat{\xi}$. The intensity $\alpha = 0.0003$ is chosen as a compromise to get a number of sample points (3,876 for the entire continent in our example, see Figure 9.1) that is high enough to ensure a reliable estimation but still allows computations to be done in a reasonable time.

The most intuitive approach is to first estimate the cross-covariance function $C_{\Pi\Lambda}$ using the method of moments, see Section 3.2.6. However, the method of moments estimator has certain disadvantages, see Remark 3.2.10, making it an inappropriate choice in this context. Also the fitting of parametric models, as advised in most geostatistical applications to obtain smooth and continuous estimates, see, e.g., Montero et al. (2015), is not suitable since it seems hardly possible to find a parametric model for cross-covariance functions that can be fitted adequately for all taxa and time periods. A more appropriate alternative is to use a kernel estimator for cross-covariance functions introduced in Section 3.2.6. This kind of estimator is frequently used in spatial statistics for the estimation of mark correlation functions, see Illian et al. (2008), and has been successfully applied and interpreted in an ecological context, see, e.g., Shimatani (2002) or Ledo et al. (2011). Let $r_{est} \in (0, r_0]$ be chosen in such a way that for each $r \leq r_{est}$, the number of pairs of sampling points $u_i, u_j \in \{u_1, \ldots, u_k\}$ in $\hat{\xi}$ with approximate distance $r \approx d_{GC}(u_i, u_j)$ is sufficiently large. For example, $r_{est} = 1000$ km is a reasonable choice for all taxa considered in this chapter. According to (3.14), an estimate $\hat{C}_{\Pi\Lambda}$ of $C_{\Pi\Lambda}$ based on $\hat{\pi}(u_1), \ldots, \hat{\pi}(u_k)$ and $\hat{\lambda}(u_1), \ldots, \hat{\lambda}(u_k)$ can be computed by

$$\hat{C}_{\Pi\Lambda}(r) = \frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{k}(\hat{\pi}(u_i) - \hat{\mu}_\Pi)(\hat{\lambda}(u_j) - \hat{\mu}_\Lambda)\kappa^E\left(\frac{r - d_{GC}(u_i,u_j)}{h}\right)}{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{k}\kappa^E\left(\frac{r - d_{GC}(u_i,u_j)}{h}\right)}, \quad r \in [0, r_{est}], \tag{9.2}$$

where h is a positive bandwidth, $\hat{\mu}_\Pi$ and $\hat{\mu}_\Lambda$ are method of moment estimates of the expectations μ_Π and μ_Λ, see (3.9), and $\kappa^E : \mathbb{R} \to [0, \infty)$ denotes the one-dimensional Epanechnikov kernel defined in (3.30). Comparisons of estimates based on different bandwidths have shown that $h = 20$ km is a good choice to obtain smooth functions without eliminating important details. Using different types of kernel functions has a negligible effect on obtained results. A similar kernel estimator for covariance functions is also proposed in Hall et al. (1994), where the authors propose an additional adjustment based on Fourier transforms to ensure that estimated functions are positive semi-definite. We are not following this path, since the procedure is computationally expensive and does not really affect qualitative interpretation of cross-correlation functions.

Finally, a plug-in estimate $\hat{\rho}_{\Pi\Lambda}$ of the cross-correlation function $\rho_{\Pi\Lambda}$ is given according to (3.16), i.e.,

$$\hat{\rho}_{\Pi\Lambda}(r) = \frac{\hat{C}_{\Pi\Lambda}(r)}{\sqrt{\hat{\sigma}_\Pi^2\,\hat{\sigma}_\Lambda^2}}, \quad r \in [0, r_{est}], \tag{9.3}$$

where $\hat{\sigma}_\Pi^2$ and $\hat{\sigma}_\Lambda^2$ are estimates of the variances σ_Π^2 and σ_Λ^2. To obtain stable estimates of cross-correlation functions, it is recommended not to use the standard moment estimates here. Instead, due to $\sigma_\Pi^2 = \operatorname{cov}(\Pi(t), \Pi(t))$ for all $t \in W$, an estimate $\hat{\sigma}_\Pi^2$ can be computed according to (9.2) for $r = 0$ with $\hat{\pi}(u_1), \ldots, \hat{\pi}(u_k)$ and $\hat{\mu}_\Pi$ instead of $\hat{\lambda}(u_1), \ldots, \hat{\lambda}(u_k)$ and $\hat{\mu}_\Lambda$, compare to (3.13). An estimate $\hat{\sigma}_\Lambda^2$ of the variance σ_Λ^2 is determined analogously. The value $\hat{\rho}_{\Pi\Lambda}(r)$ for any $r \in [0, r_{est}]$ describes the estimated correlation of the vegetation intensity and the population intensity at two arbitrary locations in the estimated taxon range $\hat{\xi}$ with a distance of r km.

9.2. Cross-correlation functions of changes in vegetation and population with temporal lag

An even more interesting question when analyzing prehistoric vegetation and population intensities is how native populations responded to environmental changes and, vice versa, how vegetation composition changed due to paleodemographic developments. In this context, it seems plausible that a change in vegetation intensity does not necessarily cause a response in population intensity immediately but a few hundred years later (or that a change in vegetation is influenced by a demographic change some hundred years earlier). For that purpose, we estimate and interpret cross-correlation functions of 500-year changes in vegetation and population intensity maps with temporal lag. We fix one of the 10 taxa considered in Section 9.1, a year y from the sequence 1000 BP, 1100 BP, $\ldots$, 12500 BP, and a temporal lag $\tau \in \{-1000, -900, \ldots, 900, 1000\}$ in years. Let $W^{(y)} \subset \mathbb{R}^2$ describe the compact set of all locations in North America that are not permanently covered by ice in the years $y - 250$ and $y + 250$, i.e., $W^{(y)}$ is the intersection of the geographical domains that are considered when estimating vegetation intensity maps for those two years. Similarly, $W^{(y+\tau)} \subset \mathbb{R}^2$ contains all locations on the North American continent that are not covered by ice in the years $y + \tau - 250$ and $y + \tau + 250$. In order to provide a consistent domain for the random fields to be compared, we furthermore introduce the intersection $\tilde{W} = W^{(y)} \cap W^{(y+\tau)}$. Additionally, let $\tilde{r}_0 > 0$ be the maximum distance such that for each $r \leq \tilde{r}_0$ there is at least one pair s, t of locations in $\tilde{W}$ with $d_{GC}(s,t) = r$, compare to (3.1). By $\hat{\pi}^{(y)}(t)$ we denote the 500-year change in estimated vegetation intensities at location $t \in \tilde{W}$ between the years $y + 250$ and $y - 250$, which is computed based on the corresponding estimated vegetation intensity maps obtained according to (8.3). Analogously, $\hat{\lambda}^{(y+\tau)}(t)$ denotes the 500-year change in estimated population intensities at $t \in \tilde{W}$ between the years $y + \tau + 250$ and $y + \tau - 250$ (or, to be more precise, between the 500-year intervals $[y + \tau, y + \tau + 500]$ and $[y + \tau - 500, y + \tau]$, whose centers are given by $y + \tau + 250$ and $y + \tau - 250$), compare to (7.3). For the estimation of cross-correlation functions we only consider locations that fall into the intersection $\tilde{\xi}$ of $\tilde{W}$ with the estimated taxon ranges that correspond to the years $y + 250$ and $y - 250$. By $\tilde{u}_1, \ldots, \tilde{u}_{\tilde{k}}$ with $\tilde{k} \in \mathbb{N}$ we denote those sample points obtained from the realization of a stationary Poisson point process depicted in Figure 9.1 that fall into $\tilde{\xi}$.

Again, we suppose that $\tilde{\xi}$ is a realization of some random closed subset $\tilde{\Xi}$ of $\tilde{W}$, see Section 3.3.6. Furthermore, we interpret the fields $\{\hat{\pi}^{(y)}(t), t \in \tilde{\xi}\}$ and $\{\hat{\lambda}^{(y+\tau)}(t), t \in \tilde{\xi}\}$ as realizations of random fields $\Pi^{(y)} = \{\Pi^{(y)}(t), t \in \tilde{W}\}$ and $\Lambda^{(y+\tau)} = \{\Lambda^{(y+\tau)}(t), t \in \tilde{W}\}$

restricted to $\tilde{\Xi}$, where $\Pi^{(y)}(t)$ and $\Lambda^{(y+\tau)}(t)$ take values in $\mathbb{R}$ for each $t \in \tilde{W}$. Next, we again suppose that the random fields $\Pi^{(y)}$ and $\Lambda^{(y+\tau)}$ are jointly second-order motion-invariant, see Section 3.2.4. Accordingly, the motion-invariant cross-covariance function $C_{\Pi^{(y)}\Lambda^{(y+\tau)}} : [0, \tilde{r}_0] \to \mathbb{R}$ of $\Pi^{(y)}$ and $\Lambda^{(y+\tau)}$ is defined by

$$C_{\Pi^{(y)}\Lambda^{(y+\tau)}}(r) = \operatorname{cov}\left(\Pi^{(y)}(s), \Lambda^{(y+\tau)}(t)\right), \quad s, t \in \tilde{W} \text{ such that } d_{GC}(s,t) = r.$$

The corresponding (motion-invariant) cross-correlation function $\rho_{\Pi^{(y)}\Lambda^{(y+\tau)}} : [0, \tilde{r}_0] \to [-1, 1]$ is represented as

$$\rho_{\Pi^{(y)}\Lambda^{(y+\tau)}}(r) = \frac{C_{\Pi^{(y)}\Lambda^{(y+\tau)}}(r)}{\sqrt{\sigma^2_{\Pi^{(y)}}\,\sigma^2_{\Lambda^{(y+\tau)}}}}, \quad r \in [0, \tilde{r}_0],$$

where $\sigma^2_{\Pi^{(y)}} = \operatorname{var}\Pi^{(y)}(t)$ and $\sigma^2_{\Lambda^{(y+\tau)}} = \operatorname{var}\Lambda^{(y+\tau)}(t)$ for any $t \in \tilde{W}$. For the estimation of $\rho_{\Pi^{(y)}\Lambda^{(y+\tau)}}$, we first consider a distance $\tilde{r}_{est} \in (0, \tilde{r}_0]$ such that for each $r \leq \tilde{r}_{est}$ the number of pairs of sampling points $\tilde{u}_i, \tilde{u}_j \in \{\tilde{u}_1, \ldots, \tilde{u}_{\tilde{k}}\}$ in $\tilde{\xi}$ with approximate distance $r \approx d_{GC}(\tilde{u}_i, \tilde{u}_j)$ is reasonably large (e.g., $\tilde{r}_{est} = 1000$ km). Finally, estimates $\hat{\rho}_{\Pi^{(y)}\Lambda^{(y+\tau)}}(r)$ of $\rho_{\Pi^{(y)}\Lambda^{(y+\tau)}}(r)$ for $r \in [0, \tilde{r}_{est}]$ can be computed using the kernel estimators given in Section 9.1 based on $\hat{\pi}^{(y)}(\tilde{u}_1), \ldots, \hat{\pi}^{(y)}(\tilde{u}_{\tilde{k}})$ and $\hat{\lambda}^{(y+\tau)}(\tilde{u}_1), \ldots, \hat{\lambda}^{(y+\tau)}(\tilde{u}_{\tilde{k}})$ instead of $\hat{\pi}(u_1), \ldots, \hat{\pi}(u_k)$ and $\hat{\lambda}(u_1), \ldots, \hat{\lambda}(u_k)$, see (9.2) and (9.3).

Note that, in a similar way, the cross-correlation function of $\Pi^{(y)}$ and $\Pi^{(y+\tau)}$ or that of $\Lambda^{(y)}$ and $\Lambda^{(y+\tau)}$ could be estimated in order to analyze the persistence of trends in the temporal development of vegetation or population intensities. This would provide additional insight into prehistoric demography and vegetation change although we have to mention that some correlation is artificially created by the applied temporal smoothing methods. However, this goes beyond the scope of this thesis.

9.3. Nonparametric estimation of pointwise confidence bands

When interpreting estimated cross-correlation functions, it usually is of great interest to know which results can be considered as significant and which not. For example, if the applied estimator has a high variance, then it is possible that a clearly positive or negative cross-correlation is obtained although there is no relationship between the underlying random fields. In order to determine which values of estimated cross-correlation functions can be considered as significantly different from zero, we compute pointwise confidence bands using nonparametric resampling methods. Several approaches are considered in the following: a subsampling method with different parameters and a bootstrap method, see, e.g., Politis et al. (1999) or Chernick and

LaBudde (2011). We describe how the suggested methods are applied to estimated cross-correlations of vegetation and population intensity maps obtained in Section 9.1. An application to cross-correlation functions of vegetation and population changes, see Section 9.2, works analogously. Let $\{\hat{\rho}_{\Pi\Lambda}(r), r \in [0, r_{est}]\}$ be an estimated cross-correlation function for some fixed year and taxon (with estimated range $\hat{\xi}$) obtained according to (9.3) based on estimated vegetation intensities $\hat{\pi}(u_1), \ldots, \hat{\pi}(u_k)$ and population intensities $\hat{\lambda}(u_1), \ldots, \hat{\lambda}(u_k)$ at sampling points $u_1, \ldots, u_k \in \hat{\xi}$. Resampling methods involve that the values of the cross-correlation function $\rho_{\Pi\Lambda}$ are re-estimated using a modified data basis. In the subsampling approach with parameter $\beta \in (0, 1)$, a random Bernoulli experiment with success probability β is performed for each sample point $u \in \{u_1, \ldots, u_k\}$ (independently for different sample points) to decide whether the data pair $(\hat{\pi}(u), \hat{\lambda}(u))$ is accepted or rejected. Based on all accepted data pairs, another estimate $\{\hat{\rho}^{(1)}_{\Pi\Lambda}(r), r \in [0, r_{est}]\}$ of $\{\rho_{\Pi\Lambda}(r), r \in [0, r_{est}]\}$ is computed according to (9.3). In contrast, when applying the bootstrap approach, the estimate $\{\hat{\rho}^{(1)}_{\Pi\Lambda}(r), r \in [0, r_{est}]\}$ is determined based on $\hat{\pi}(u'_1), \ldots, \hat{\pi}(u'_k)$ and $\hat{\lambda}(u'_1), \ldots, \hat{\lambda}(u'_k)$, where the k sample points $u'_1, \ldots, u'_k$ are drawn randomly with replacement from the set $\{u_1, \ldots, u_k\}$ of original sample points. The difference between both methods can be interpreted as follows. In the subsampling approach, the cross-correlation function is re-estimated based on a reduced set of sampling points (which means a reduced data set), i.e., it is analyzed how estimates of cross-correlation functions change when it is assumed that some of the data are overrepresented or erroneous and are thus dropped. In the bootstrap approach, however, we do not only drop some of the sample points (and the corresponding data) but also provide the estimated intensities at some other sample points with a higher weight to consider the case that those values might be underrepresented in the data. Note that in both approaches, the spatial correlation structure of the data is not violated since estimated vegetation and population intensities (at accepted or randomly drawn sampling points) are still associated with the same geographical locations as before. The considered resampling procedure (subsampling or bootstrap) is repeated 5,000 times, which results in a sample $\hat{\rho}^{(1)}_{\Pi\Lambda}(r), \ldots, \hat{\rho}^{(5000)}_{\Pi\Lambda}(r)$ of cross-correlation estimates for each $r \in [0, r_{est}]$. Based on this sample, a confidence interval $[\theta_{(\gamma)}(r), \theta^{(\gamma)}(r)]$ of level $\gamma \in (0, 1)$ for $\hat{\rho}_{\Pi\Lambda}(r)$ is constructed, where $\theta_{(\gamma)}(r)$ denotes the empirical $(1-\gamma)/2$ quantile and $\theta^{(\gamma)}(r)$ is the empirical $1-(1-\gamma)/2$ quantile of the sample $\hat{\rho}^{(1)}_{\Pi\Lambda}(r), \ldots, \hat{\rho}^{(5000)}_{\Pi\Lambda}(r)$. Consequently, the confidence interval $[\theta_{(\gamma)}(r), \theta^{(\gamma)}(r)]$ contains $\gamma \cdot 100\%$ of the estimates $\hat{\rho}^{(1)}_{\Pi\Lambda}(r), \ldots, \hat{\rho}^{(5000)}_{\Pi\Lambda}(r)$. An estimated cross-correlation $\hat{\rho}_{\Pi\Lambda}(r)$ is considered to be significantly different from zero, if $0 \notin [\theta_{(\gamma)}(r), \theta^{(\gamma)}(r)]$, where the level γ is typically chosen as $\gamma = 0.95$ or $\gamma = 0.99$. Finally, the functions $\{\theta_{(\gamma)}(r), r \in [0, r_{est}]\}$ and $\{\theta^{(\gamma)}(r), r \in [0, r_{est}]\}$ describe the (lower and upper) boundaries of pointwise confidence bands of level γ for the estimated cross-correlation function $\{\hat{\rho}_{\Pi\Lambda}(r), r \in [0, r_{est}]\}$.

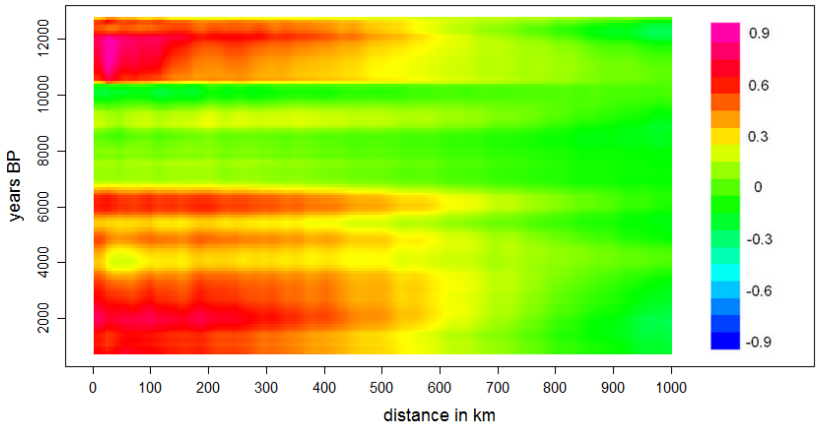


Figure 9.2.: Estimated cross-correlation functions for vegetation intensity maps of *Quercus* and population intensity maps between 12750 BP and 750 BP.

9.4. Discussion of correlation results for Quercus

To conclude this chapter, we briefly present the results of the performed cross-correlation analysis, where we again focus on the history of *Quercus* (oak). A more detailed discussion for further taxa goes beyond the scope of this thesis and is provided in Gajewski et al. (2017). In Figure 9.2, estimated cross-correlation functions $\hat{\rho}_{\Pi\Lambda}$ obtained according to (9.3) for all years $y \in \{750, 850, \ldots, 12750\}$ are depicted, where warm colors indicate positive and cold colors negative values. Prior to approximately 10500 BP, intensities of both *Quercus* and population were high in Florida and decreased towards the north and west, compare to Sections 7.3 and 8.4, which leads to high estimated cross-correlations. Between 10500 BP and 6800 BP, the range of *Quercus* expanded and moved northwards, while populations showed a complex pattern of changes resulting in cross-correlations that are close to zero. For example, in Texas relatively low values of *Quercus* and high population intensities are observed, whereas in Florida *Quercus* remained stable but population numbers decreased. This was a time period of quite some variability in the climate, see Viau et al. (2006), although incompleteness of the CARD cannot be ruled out. During the period between ca. 6800 BP and 3700 BP, the Late Archaic cultural period, cross-correlations started to increase again but also showed some fluctuations. The distribution of *Quercus* did not change much during this time but population gradually increased in the central regions of the eastern US, where *Quercus* had its maximum intensities. However, population numbers also were subject to certain fluctuations causing cross-correlations to increase and decrease several times during this period. Over the past 3600 years, the Woodland Period, cross-correlations were uniformly high. Populations decreased in New England but greatly increased in the area of maximum *Quercus* abundance in the eastern US. The large values of estimated cross-correlation functions in the late Holocene indicate that forests with a high abundance of *Quercus* provide optimal conditions for population growth during this time. However, potential issues with incompleteness of the CARD (particularly in the south, see Section 7.1), or the low density of pollen samples in the southeastern US, see Figure 8.3, may have contributed to obtained cross-correlations as well.

Next, estimated cross-correlation functions $\hat{\rho}_{\Pi^{(y)}\Lambda^{(y+\tau)}}$ of 500-year changes in *Quercus* abundance and population intensity as described in Section 9.2 are analyzed. Figure 9.3 illustrates the values of cross-correlation functions for $y \in \{1000, 1100, \ldots, 12500\}$ and three different temporal lags, where y describes the midpoint of the 500-year interval in which changes of vegetation intensities are computed. However, it is desirable to summarize cross-correlations for all temporal lags in one figure to provide a compact overview of spatio-temporal relationships between vegetation and population changes. For that purpose, we compute for each year $y \in \{1000, 1100, \ldots, 12500\}$ and each lag $\tau \in \{-1000, -900, \ldots, 900, 1000\}$ the mean value of the estimated cross-correlations $\{\hat{\rho}_{\Pi^{(y)}\Lambda^{(y+\tau)}}(r), r = 30, 35, 40, \ldots, 200 \text{ km}\}$ (the boundaries are chosen to reflect local associations only and to avoid unstable estimates for distances close to zero) and summarize computed means in a matrix representation, see Figure 9.4. Values on the bold diagonal line correspond to the temporal lag $\tau = 0$ (compare to Figure 9.3, center), whereas cross-correlations above the diagonal line correspond to a negative temporal lag and values below the line to a positive lag. Several periods of positive and negative cross-correlations between changes in *Quercus* and population are found, some of which are briefly discussed in the following.

Large positive values of cross-correlations are observed for changes in *Quercus* between 11000 BP and 9500 BP and changes in population in the period between 11300 BP and 10000 BP for all positive and small negative temporal lags. This indicates a strong relationship between demographic and environmental changes over more than 1000 years. For example, both vegetation and population intensities were decreasing in Florida and increasing further north (especially *Quercus*) during this period. This was a time of rapid warming in North America, see Viau et al. (2006), where native populations relied on a diverse set of resources, including hunting different animals and gathering a variety of food, see Fagan (2000). The increased diversity of forests in which *Quercus* typically occurs would have provided a suitable habitat to enable the increase of human populations. The period with low to medium negative cross-correlations between 7500 BP and 6500 BP is a time in which both vegetation intensities and population intensities showed a complex series of developments. Major changes in *Quercus* (due to a moister climate) are observed in the west, whereas changes in population occurred in different regions across the range of *Quercus*. Also a decrease in *Quercus* abundance in Florida and the southeast (combined with increasing populations in these regions) has probably caused negative cross-correlations. An increase in sedentism and the use of diverse resources by native populations could have enabled adaptation to numerous

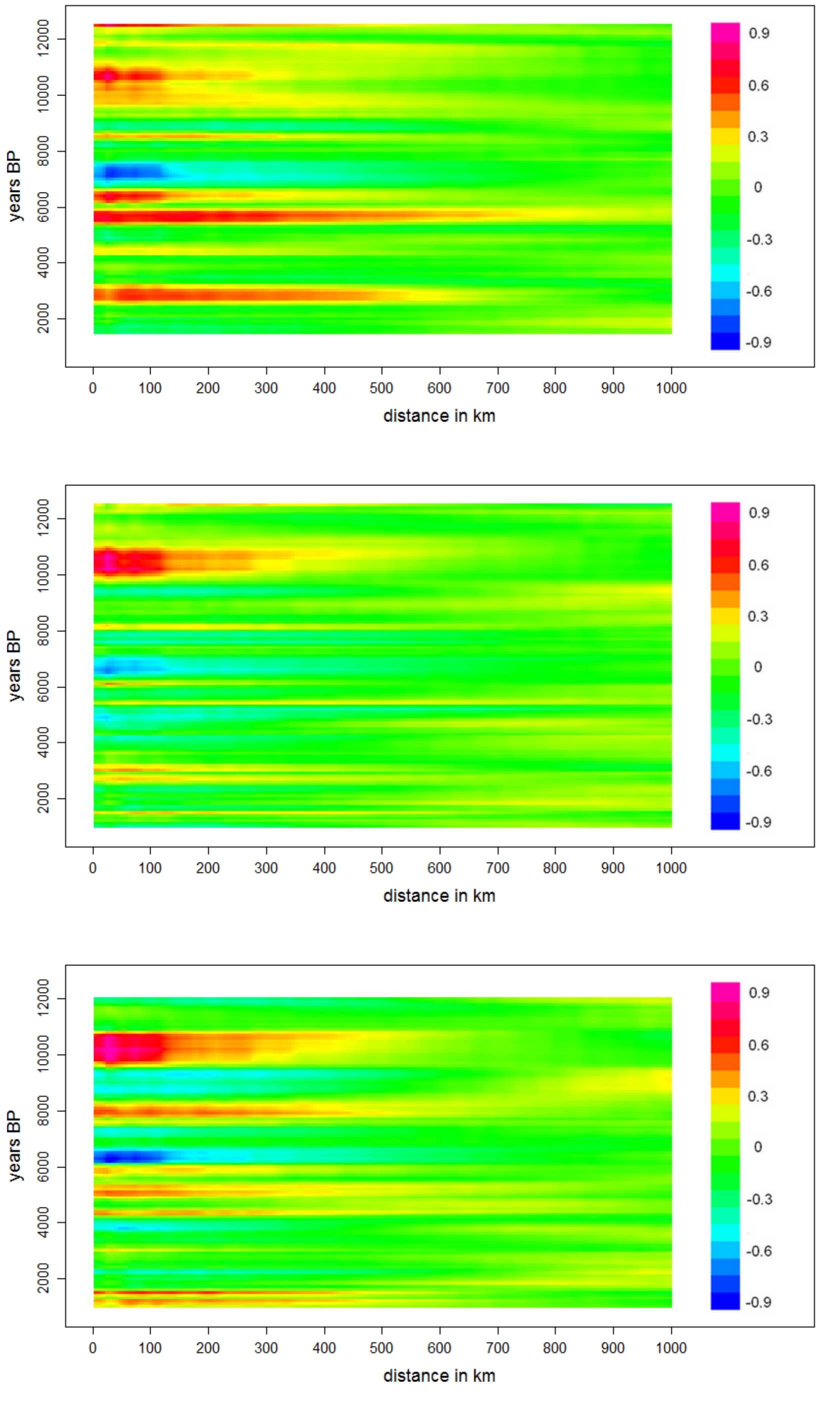


Figure 9.3.: Estimated cross-correlation functions for 500-year changes in vegetation intensity maps of *Quercus* and population intensity maps with three temporal lags: $\tau = -500$ years (top), $\tau = 0$ years (center), and $\tau = 500$ years (bottom). Values on the ordinate correspond to the midpoints of the 500-year change intervals for vegetation intensities.

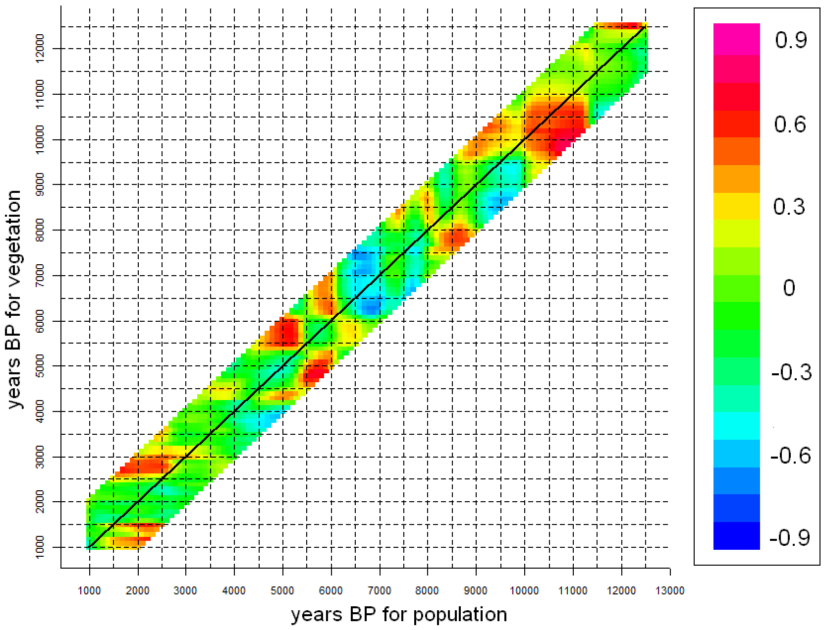


Figure 9.4.: Mean cross-correlations (of distances between 30 and 200 km) for 500-year changes in vegetation intensity maps of *Quercus* and population intensity maps with different temporal lags. Values on the abscissa and ordinate correspond to the midpoints of the 500-year change intervals for population intensities and vegetation intensities, respectively.

environments, which further contributes to a lower association with *Quercus* abundance, see Fagan (2000). Moreover, in Delcourt and Delcourt (2004) it is speculated that in some regions prehistoric people thinned tree populations to increase acorn yields, which could explain negative correlations, too. During the most recent 3000 years, agriculture and more sedentary lifestyles became increasingly established in the eastern US and southeastern Canada, with large cities in some areas and increasing impact on the environment, see, e.g., Fagan (2000), Delcourt and Delcourt (2004) or Munoz and Gajewski (2010). Agriculture (which is associated to increasing populations) and *Quercus* abundance would both be favored in areas with optimal environmental conditions. This could explain positive cross-correlations of changes in *Quercus* between 3000 BP and 2500 BP and population between 2500 BP and 1500 BP as well as of changes in *Quercus* between 1500 BP and 1000 BP and population between 2500 BP and 1500 BP.

To conclude presentation of correlation results, we discuss pointwise confidence bands for cross-correlation functions as described in Section 9.3. Due to their time-consuming computation, confidence bands are only provided every 1000 years for cross-correlation

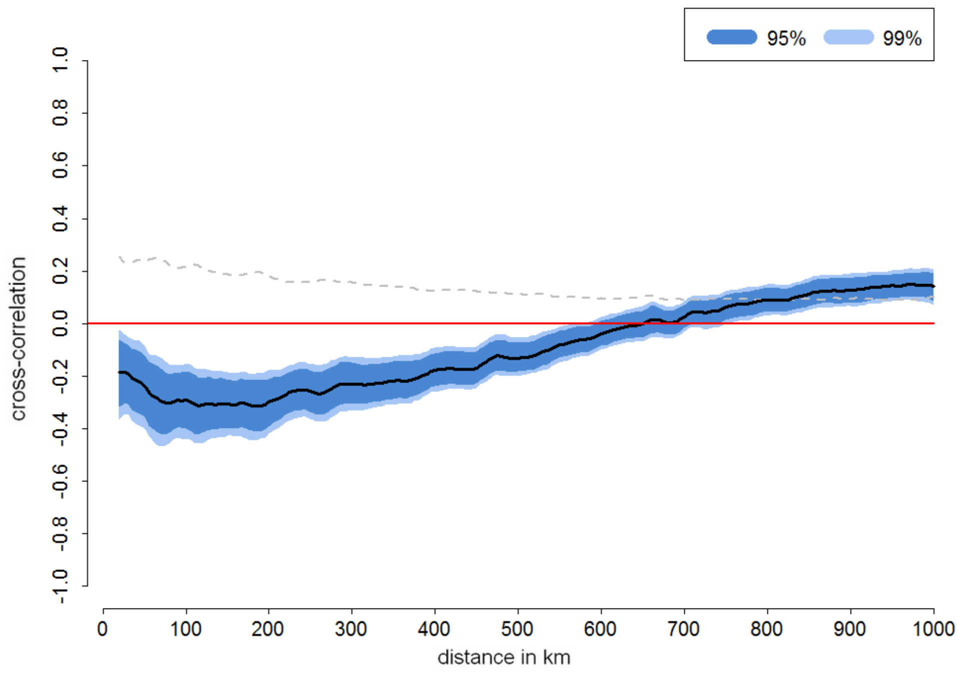


Figure 9.5.: Estimated cross-correlation function for 500-year changes in vegetation intensities of *Quercus* and population intensities between 2550 BP and 2050 BP together with pointwise confidence bands of levels 0.95 and 0.99 computed based on the bootstrap method. The gray dashed line indicates the difference between the upper and the lower bound of level 0.95.

functions of vegetation and population intensity maps as well as for cross-correlations of changes in vegetation and population with temporal lag $\tau = 0$. Furthermore, both the bootstrap method and the subsampling method (with different parameters $\beta_1 = 0.25$, $\beta_2 = 0.5$, and $\beta_3 = 0.75$) are used. Figure 9.5 illustrates pointwise confidence bands for changes in vegetation intensities (of *Quercus*) and population intensities between 2550 BP and 2050 BP, which are computed using the bootstrap method. We observe that cross-correlations for distances up to 550 km can be considered to be significantly different from zero at level 0.99. The subsampling approach with parameter $\beta_2 = 0.5$ leads to bands that are almost identical to those depicted in Figure 9.5. Using subsampling with $\beta_1 = 0.25$ or $\beta_3 = 0.75$ results in wider or narrower bands, respectively. Finally, we aim to give a general recommendation on how large estimated cross-correlations of changes in *Quercus* and population (with temporal lag $\tau = 0$) should be in order to be considered as significantly different from zero for all time periods. For each computed confidence band, we determine half of the maximum difference between the upper and the lower bound of level 0.95 for all distances between 30 and 200 km and plot these values as a function of the

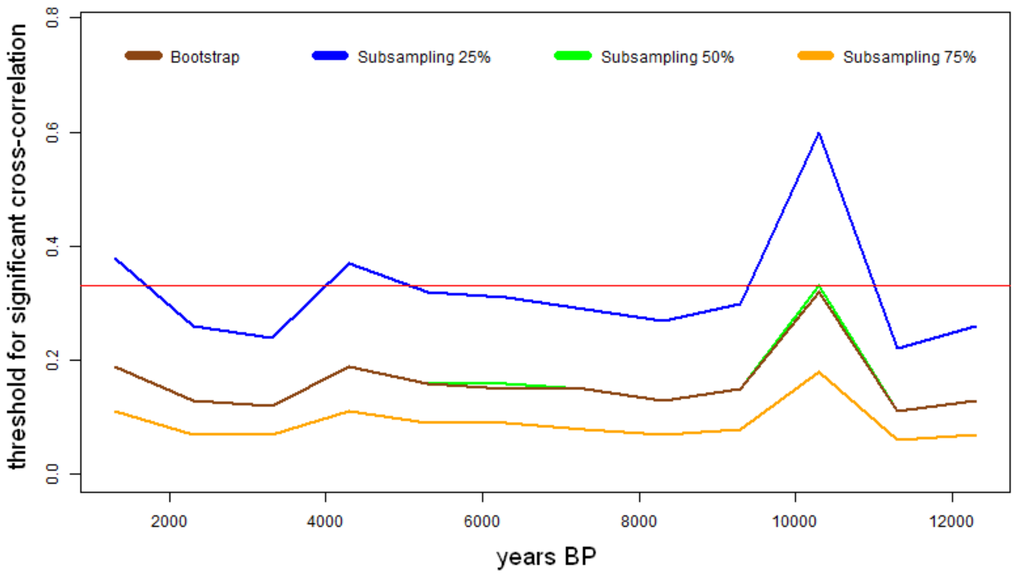


Figure 9.6.: Graphs showing the significance of estimated cross-correlation functions for 500-year changes in vegetation intensities of *Quercus* and population intensities based on pointwise confidence bands of level 0.95 for four different methods. For each computed confidence band, half of the maximum difference between the upper and the lower bound among distances between 30 and 200 km is illustrated as a function of the corresponding year.

corresponding year, see Figure 9.6. For example, the maxima of the brown and the green graph indicate that when relying on the bootstrap approach or the subsampling method with parameter $\beta_2 = 0.5$ (which is a reasonably conservative choice), then all cross-correlations (for all considered time periods) that are greater than 0.35 (or smaller than -0.35) can be considered to be significantly different from zero. Note that this approach assumes confidence bands to be symmetric, which is the case (at least approximately) for the vast majority of estimated cross-correlation functions. When interpreting cross-correlations for vegetation intensities of *Quercus* and population intensities as illustrated in Figure 9.2 (instead of changes), similar plots can be derived, which reveal that all (absolute) values greater than 0.3 are significantly different from zero.

10. Conclusions

In the present thesis, we discuss various stochastic models, statistical methods, and simulation algorithms to address open questions in PWP, archeology, and paleoecology. In Chapter 4, a stochastic model for precipitation cells is proposed to enable the computation of area probabilities for the occurrence of precipitation. In the presented approach, precipitation cells are described using a germ-grain model with circular grains, which is based on a non-stationary Cox point process. This model is extended in Chapter 5 to allow for the estimation of area probabilities for the occurrence of higher precipitation amounts. For that purpose, a randomly scaled response function is assigned to each precipitation cell and the summed response functions are interpreted as random precipitation amounts. The most important model characteristics (intensities of precipitation cells, cell radius, expectations and variances of random scaling variables) are computed algorithmically based on point probabilities (separately for each forecast period), i.e., no precipitation data are required for model fitting. Estimators are provided to determine area probabilities based on repeated simulation of the proposed model. A comparison of estimated area probabilities with radar data shows a close correspondence. For thresholds up to 3 mm, we receive reasonable BSSs and correlation coefficients for almost all test areas, which in many cases are even higher than the scores of the underlying point probabilities. In general, biases are close to zero but also show some deviations for few test areas. For higher thresholds of 5 mm or more, forecast verification shows less significant results. In particular, BSSs and correlation coefficients are lower than for smaller thresholds and area probabilities seem to be marginally underestimated. However, we still obtain positive BSSs and correlation coefficients for a majority of test areas, which indicates that estimated probabilities have a higher quality than the climate mean. The analysis of computed RPSSs also confirms these results. We get reasonably high values for both point and area probabilities, which demonstrates a strong relationship between estimated probabilities and radar data. This is also indicated by computed reliability diagrams of area probabilities for the occurrence of precipitation, which show a particularly nice performance, too. However, we find that all considered verification scores of area probabilities clearly depend on the scores of underlying point probabilities. Thus, reliable, unbiased data are crucial for the applicability of the presented method. Although it remains unclear whether the lower quality of area probabilities for higher thresholds is caused by the underlying data or the stochastic model, we will consider the derivation of more precise area probabilities for rare precipitation events as a major goal of future research. In principle, the combined model of precipitation cells and precipitation amounts can also be applied to determine further characteristics that are relevant for the issuing of weather warnings. For example, it seems possible to use the model for the estimation of the expected cumulated precipitation amount that occurs in the drainage area of a river to assess flood risks. However, this is beyond the scope of this thesis and further research is required to assess the applicability of the precipitation model to such problems.

A further generalization of the model for precipitation cells is provided in Chapter 6, where we propose a representation of thunderstorm cells based on spatial cluster processes. In contrast to the previous approaches, which solely rely on available point probabilities, the cluster model also requires thunderstorm records of past periods to statistically determine all model characteristics. Moreover, formulas for the computation of point and area probabilities are derived and a forecast verification is performed using thunderstorm data provided by DWD. It turns out that this more complex model is able to produce reliable area probabilities for the occurrence of thunderstorms, which have a higher forecast quality (according to computed verification scores such as bias, LSS, and empirical correlation coefficient) than the underlying point probabilities. The performance of forecasts can be increased even more if realizations of the model are generated conditionally on thunderstorm data from past periods, which provides a seamless combination of thunderstorm records and point probabilities from PWP for the issuing of short-term weather warnings. However, currently the model is designed for short forecast ranges only, especially an application of the conditional simulation algorithm only provides better results for ranges up to three hours ahead. Therefore, a main topic of future research could be a generalization of the proposed model with the goal of generating more realistic realizations of thunderstorms for arbitrary forecast ranges without using the conditional simulation algorithm. Possible ideas include the incorporation of elliptic clusters, spatially varying cluster parameters or even the integration of precipitation produced by single thunderstorm cells. Furthermore, an adaption of the proposed conditional simulation algorithm to the model for precipitation cells is currently under preparation.

In summary, the model-based methods proposed in Part II of this thesis are expected to play a central role when addressing a specific responsibility of DWD, which is the derivation and dissemination of area probabilities for potentially dangerous weather events for customer-specific areas using semi-automated warning systems. For the first time, spatial stochastic models of precipitation amounts and thunderstorm cells are provided, which allow for the estimation of precise area probabilities while fulfilling important requirements for application in modern weather prediction. The proposed models do neither assume spatial nor temporal stationarity and all model characteristics are determined automatically from point forecasts or observation data from past periods. This, together with the reasonable computation time, makes the considered approaches suitable for an operational derivation of weather forecasts for the territory of Germany.

However, it remains to be investigated how the presented methods perform for areas with sizes and shapes varying from those investigated in our example of application or in regions with different climate conditions than Germany. In general, an adaption of the proposed methodology to further weather events seems possible although a verification of area forecasts is difficult due to the lack of spatially inclusive and comprehensive high-resolution observation data for most meteorological variables.

Part III of the present thesis addresses the statistical estimation of prehistoric population and vegetation intensity maps for North America with the goal of analyzing interrelationships between demographic developments and environmental changes over the past 13000 years. At first, Chapter 7 describes a nonparametric statistical methodology to the reconstruction of population intensities for a sequence of 500-year intervals, which takes into account errors occurring due to inhomogeneous sampling strategies, taphonomic loss, and boundary effects. The resulting maps show temporally distinct dynamic patterns of paleodemographic trends that correspond well to independent archaeological, ethnohistoric, and genetic evidence from literature. Several sensitivity and robustness analyses reveal that most modifications of the method and the database only marginally influence the obtained results. For the broad population patterns, the effect of applying an adaptive smoothing method is just as small as that of randomly dropping up to 50 % of the underlying radiocarbon dates. The only significant effect is observed when modifying the globally constant bandwidth. While in this case a large effect on the spatial scale of population events is visible, affected regions still remain the same. As the underlying radiocarbon database, the CARD, is currently extended to other continents, a global estimation of prehistoric population intensities could be an interesting topic of future research. Furthermore, such results have implications for hypothesizing and testing human migration routes as well as the relative influence of human populations on the evolution of ecosystems. In Chapter 8, a similar statistical method is designed to estimate comparable vegetation intensity maps for 10 plant taxa based on pollen data obtained from a comprehensive paleoecological database. We briefly describe and interpret results for *Quercus* (oak), which correspond well with existing estimates from literature.

Finally, Chapter 9 provides a simple approach to analyze statistical cross-correlations of (changes in) population and vegetation intensity maps for various spatial and temporal lags. Furthermore, the computation of pointwise confidence bands based on nonparametric resampling methods is considered to determine the significance of obtained correlation results. This analysis is motivated by the fact that relationships between demographic and environmental changes are complex and causal associations can go in both directions. In addition, the human use of natural resources as well as influences of populations on the environment have changed through time as cultures evolved. The proposed approach is again illustrated using the example of *Quercus*. By estimating and interpreting cross-correlations between 500-year changes in population intensities and the relative abundance of *Quercus* in eastern North America, we could identify times in the past with both positive and negative associations. However, a more detailed analysis of the estimated maps reveals times and locations where individual data points seem to have undue influence on obtained cross-correlations, especially in regions of low sampling intensities. Thus, future work is planned to refine the method, e.g., by determining minimum site densities that are needed to provide more reliable estimates or by better identifying overly influential points and outliers. Local analyses with smaller smoothing bandwidths may also be of great interest. Furthermore, a companion study is prepared, see Gajewski et al. (2017), where results for the other taxa are described and interpreted, too. Since the community ecology of the considered tree species is well understood and human use of the various taxa has also been studied several times in the literature, consistencies between the taxa could help to better identify significant signals in obtained correlation results. In conclusion, the considered approach is a first attempt to quantify relationships between prehistoric demographic changes and environmental conditions on a continental scale and is thus expected to contribute to the understanding of human-environment interactions in North America during the Holocene. Moreover, the proposed methodology has a wide range of further potential applications in geographic and environmental sciences, where large and complex databases consisting of repeated measurements at a spatial system of sites are compiled.

Bibliography

Adler, R. J. (2010). *The Geometry of Random Fields.* Society for Industrial and Applied Mathematics, Philadelphia.

Baddeley, A., Bárány, I., Schneider, R., and Weil, W. (2007). *Stochastic Geometry.* Springer, Berlin.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., and Reinhardt, T. (2011). Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Mon. Weather Rev.*, 139:3887–3905.

Banerjee, S. (2005). On geodetic distance computations in spatial modeling. *Biom.*, 61:617–625.

Beckers, F. and Bogaert, P. (1998). Nonstationarity of the mean and unbiased variogram estimation: Extension of weighted least-squares method. *Math. Geol.*, 30:499–521.

Behrends, E. (2000). *Introduction to Markov Chains: With Special Emphasis on Rapid Mixing.* Vieweg, Braunschweig.

Benedetti, R. (2010). Scoring rules for forecast verification. *Mon. Weather Rev.*, 138:203–211.

Beneš, V. and Rataj, J. (2004). *Stochastic Geometry: Selected Topics.* Kluwer Academic Publishers, Boston.

Birks, H. J. B. and Birks, H. (1980). *Quaternary Palaeoecology.* Blackburn Press, Caldwell.

Blaauw, M. and Christen, J. A. (2011). Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Anal.*, 6:457–474.

Bonatto, S. L. and Salzano, F. M. (1997). A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. *Proc. Natl. Acad. Sci.*, 94:1866–1871.

Brereton, T. J., Stenzel, O., Baumeier, B., Andrienko, D., Schmidt, V., and Kroese, D. P. (2014). Efficient simulation of Markov chains using segmentation. *Methodol. Comput. Appl. Probab.*, 16:465–484.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, 78:1–3.

Bronk Ramsey, C. (2008). Deposition models for chronological records. *Quat. Sci. Rev.*, 27:42–60.

Burton, A., Fowler, H. J., Kilsby, C. G., and O'Connell, P. E. (2010). A stochastic model for the spatial-temporal simulation of nonhomogeneous rainfall occurrence and amounts. *Water Resour. Res.*, 46:W11501.

Burton, A., Kilsby, C. G., Fowler, H. J., Cowpertwait, P. S. P., and O'Connell, P. E. (2008). RainSim: A spatial–temporal stochastic rainfall modelling system. *Environ. Modell. Softw.*, 23:1356–1369.

Butcher, J. C. (2016). *Numerical Methods of Ordinary Differential Equations*, 3rd edition. J. Wiley & Sons, Chichester.

Campbell, N. R. (1909). The study of discontinuous phenomena. *Proc. Cambridge Philos. Soc.*, 15:117–136.

Chandler, R. E. (1997). A spectral method for estimating parameters in rainfall models. *Bernoulli*, 3:301–322.

Chaput, M. A., Kriesche, B., Betts, M., Martindale, A., Kulik, R., Schmidt, V., and Gajewski, K. (2015). Spatiotemporal distribution of Holocene populations in North America. *Proc. Natl. Acad. Sci.*, 112:12127–12132.

Chatterjee, S. and Hadi, A. S. (2012). *Regression Analysis by Example*, 5th edition. J. Wiley & Sons, Hoboken.

Chernick, M. R. and LaBudde, R. A. (2011). *An Introduction to Bootstrap Methods with Applications to R*. J. Wiley & Sons, Hoboken.

Chilès, J.-P. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*, 2nd edition. J. Wiley & Sons, New York.

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic Geometry and its Applications*, 3rd edition. J. Wiley & Sons, Chichester.

Choi, I., Li, B., and Wang, X. (2013). Nonparametric estimation of spatial and space-time covariance function. *J. Agric. Biol. Environ. Stat.*, 18:611–630.

Chorin, A. J. and Marsden, J. E. (1993). *A Mathematical Introduction to Fluid Dynamics*, 3rd edition. Springer, New York.

Christiansen, M. C., Hirsch, C., and Schmidt, V. (2014). Prediction of regionalized car insurance risks based on control variates. *Stat. Risk Model.*, 31:163–181.

Coiffier, J. (2011). *Fundamentals of Numerical Weather Prediction.* Cambridge University Press, New York.

Collard, M., Edinborough, K., Shennan, S., and Thomas, M. G. (2010). Radiocarbon evidence indicates that migrants introduced farming to Britain. *J. Archaeol. Sci.*, 37:866–870.

Cowpertwait, P. S. P. (1995). A generalized spatial-temporal model of rainfall based on a clustered point process. *Proc. R. Soc. Lond. A*, 450:163–175.

Cowpertwait, P. S. P., Isham, V. S., and Onof, C. J. (2007). Point process models of rainfall: Developments for fine-scale structure. *Proc. R. Soc. A*, 463:2569–2587.

Cowpertwait, P. S. P., Xie, G., Isham, V. S., Onof, C. J., and Walsh, D. C. I. (2011). A fine-scale point process model of rainfall with dependent puls depths within cells. *Hydrol. Sci. J.*, 56:1110–1117.

Cox, D. R. (1955). Some statistical models connected with series of events. *J. R. Stat. Soc. Ser. B*, 17:129–164.

Cox, D. R. and Isham, V. S. (1988). A simple spatial-temporal model of rainfall. *Proc. R. Soc. Lond. A*, 415:317–328.

Cressie, N. A. C. (1985). Fitting variogram models by weighted least squares. *Math. Geol.*, 17:563–586.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, revised edition. J. Wiley & Sons, New York.

Cressie, N. A. C. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data.* J. Wiley & Sons, Hoboken.

Daan, H. (1985). Sensitivity of verification scores to the classification of the predictand. *Mon. Weather Rev.*, 113:1384–1392.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods*, 2nd edition. Springer, New York.

Daley, D. J. and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes, Volume II: General Theory and Structure*, 2nd edition. Springer, New York.

Delcourt, P. and Delcourt, H. (2004). *Prehistoric Native Americans and Ecological Change.* Cambridge University Press, Cambridge.

Denevan, W. (1992). The pristine myth: The landscape of the Americas in 1492. *Ann. Assoc. Am. Geogr.*, 82:369–385.

Diggle, P. J. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, 3rd edition. Chapman & Hall/CRC, Boca Raton.

Diggle, P. J. and Ribeiro Jr., P. J. (2007). *Model-Based Geostatistics*. Springer, New York.

Dyke, A. S. (2002). The Laurentide and Innuitian ice sheets during the Last Glacial Maximum. *Quat. Sci. Rev.*, 21:9–31.

Elsner, J. B., Murnane, R. J., Jagger, T. H., and Widen, H. M. (2013). A spatial point process model for violent tornado occurrence in the US Great Plains. *Math. Geosci.*, 45:667–679.

Epstein, S. E. (1966). Point and area precipitation probabilities. *Mon. Weather Rev.*, 94:595–598.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. Second Int. Conf. Knowl. Discov. Data Min.*, pages 226–231, Portland.

Fagan, B. M. (2000). *Ancient North America*. Thames and Hudson, London.

Faulkner, B. R. (2002). Java classes for nonprocedural variogram modeling. *Comput. Geosci.*, 28:387–397.

Gaiselmann, G., Tötzke, C., Manke, I., Lehnert, W., and Schmidt, V. (2014). 3D microstructure modeling of compressed fiber-based materials. *J. Power Sources*, 257:52–64.

Gajewski, K., Kriesche, B., Chaput, M. A., Kulik, R., and Schmidt, V. (2017). Human-vegetation interactions during the Holocene in North America. *Ecol. Monogr.* (submitted).

Gajewski, K., Munoz, S., Peros, M., Viau, A. E., Morlan, R., and Betts, M. (2011). The Canadian Archaeological Radiocarbon Database (CARD): Archaeological ^{14}C dates in North America and their paleoenvironmental context. *Radiocarb.*, 53:371–394.

Gebhardt, C., Theis, S. E., Paulat, M., and Bouallègue, Z. B. (2011). Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.*, 100:168–177.

Gentle, J. E. (2003). *Random Number Generation and Monte Carlo Methods*, 2nd edition. Springer, New York.

Genton, M. G. and Gorsich, G. J. (2002). Nonparametric variogram and covariogram estimation with Fourier-Bessel matrices. *Comput. Stat. Data Anal.*, 41:47–57.

Genton, M. G. and Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. *Stat. Sci.*, 30:147–163.

Glahn, H. R. and Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.*, 11:1203–1211.

Goebel, T., Waters, M. R., and O'Rourke, D. H. (2008). The late Pleistocene dispersal of modern humans in the Americas. *Sci.*, 319:1497–1502.

Goldberg, A., Mychajliw, A. M., and Hadley, E. A. (2016). Post-invasion demography of prehistoric humans in South America. *Nat.*, 532:232–235.

Goring, S., Dawson, A., Simpson, G. L., Ram, K., Graham, R. W., Grimm, E. C., and Williams, J. W. (2015). neotoma: A programmatic interface to the Neotoma paleoecological database. *Open Quat.*, 1:1–17.

Grimm, E. C. (2008). Neotoma - an ecosystem database for the Pliocene, Pleistocene and Holocene. Technical report, Illinois State Museum.

Grimmett, G. R. and Stirzaker, D. R. (2001). *Probability and Random Processes*, 3rd edition. Oxford University Press, Oxford.

Grove, M. (2011). A spatio-temporal kernel method for mapping changes in prehistoric land-use patterns. *Archaeom.*, 53:1012–1030.

Hall, P., Fisher, N. I., and Hoffmann, B. (1994). On the nonparametric estimation of covariance functions. *Ann. Stat.*, 22:2115–2134.

Handl, L., Torbahn, L., Spettl, A., Schmidt, V., and Kwade, A. (2017). Structural analysis and tracking of micron-sized glass particles during shear deformation: A study based on time-resolved tomographic data. *Adv. Powder Technol.*, 28:1920–1929.

Härdle, W. (1991). *Smoothing Techniques*. Springer, New York.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer, Berlin.

Haslett, J. and Parnell, A. (2008). A simple monotone process with application to radiocarbon-dated depth chronologies. *J. R. Stat. Soc. Ser. C*, 57:399–418.

Hess, R., Kriesche, B., Schaumann, P., Reichert, B. K., and Schmidt, V. (2017). Area precipitation probabilities derived from point forecasts for operational weather and warning service applications. *Q. J. R. Meteorol. Soc.* (submitted).

Hirsch, C., Neuhäuser, D., Gloaguen, C., and Schmidt, V. (2015). First-passage percolation on random geometric graphs and an application to shortest-path trees. *Adv. Appl. Probab.*, 47:328–354.

Hoffecker, J. F., Powers, W. R., and Goebel, T. (1993). The colonization of Beringia and the peopling of the New World. *Sci.*, 259:46–53.

Hughes, J. P., Guttorp, P., and Charles, S. P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *Appl. Stat.*, 48:15–30.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns.* J. Wiley & Sons, Chichester.

Inness, P. M. and Dorling, S. (2013). *Operational Weather Forecasting.* J. Wiley & Sons, Chichester.

Ivanov, A. V. and Leonenko, N. N. (1989). *Statistical Analysis of Random Fields.* Kluwer Academic Publishers, Dordrecht.

Jacod, J. and Protter, P. E. (2004). *Probability Essentials*, 2nd edition. Springer, Berlin.

Jian, X., Olea, R. A., and Yu, Y.-S. (1996). Semivariogram modeling by weighted least squares. *Comput. Geosci.*, 22:387–397.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics.* Academic Press, London.

Kaczmarska, J., Isham, V. S., and Onof, C. J. (2014). Point process models for fine-resolution rainfall. *Hydrol. Sci. J.*, 59:1972–1991.

Kallenberg, O. (1986). *Random Measures*, 4th edition. Akademie-Verlag, Berlin.

Kallenberg, O. (2002). *Foundations of Modern Probability*, 2nd edition. Springer, New York.

Kaminsky, A. (2007). Parallel Java: A unified API for shared memory and cluster parallel programming in 100 % Java. In *Proc. 21st IEEE Int. Parallel Distrib. Process. Symp.*, Long Beach.

Kelly, R. L., Surovell, T. A., Shuman, B. N., and Smith, G. M. (2013). A continuous climatic impact on Holocene human population in the Rocky Mountains. *Proc. Natl. Acad. Sci.*, 110:443–447.

Kim, T. Y. and Park, J.-S. (2012). On nonparametric variogram estimation. *J. Korean Stat. Soc.*, 41:399–413.

Kingman, J. F. C. (1993). *Poisson Processes.* Oxford University Press, Oxford.

Knüpffer, K. (1996). Methodical and predictability aspects of MOS systems. In *Proc. 13th Conf. Probab. Stat. Atmos. Sci.*, pages 190–197, San Francisco.

König, D. and Schmidt, V. (1992). *Zufällige Punktprozesse.* Teubner, Stuttgart.

Koubek, A., Pawlas, Z., Brereton, T., Kriesche, B., and Schmidt, V. (2016). Testing the random field model hypothesis for random marked closed sets. *Spat. Stat.*, 16:118–136.

Kriesche, B., Chaput, M. A., Kulik, R., Gajewski, K., and Schmidt, V. (2017a). Estimation of spatio-temporal correlations of prehistoric population and vegetation in North America. *Geogr. Anal.* (submitted).

Kriesche, B., Hess, R., Reichert, B. K., and Schmidt, V. (2015a). A probabilistic approach to the prediction of area weather events, applied to precipitation. *Spat. Stat.*, 12:15–30.

Kriesche, B., Hess, R., and Schmidt, V. (2017b). A point process approach for spatial stochastic modeling of thunderstorm cells. *Probab. Math. Stat.*, 37 (in print).

Kriesche, B., Koubek, A., Pawlas, Z., Beneš, V., Hess, R., and Schmidt, V. (2015b). A model-based approach to the computation of area probabilities for precipitation exceeding a certain threshold. In *Proc. 21st Int. Congr. Model. Simul.*, pages 2103–2109, Gold Coast.

Kriesche, B., Koubek, A., Pawlas, Z., Beneš, V., Hess, R., and Schmidt, V. (2017c). On the computation of area probabilities based on a spatial stochastic model for precipitation cells and precipitation amounts. *Stoch. Environ. Res. Risk Assess.*, 31:2659–2674.

Kriesche, B., Weindl, H., Smolka, A., and Schmidt, V. (2014). Stochastic simulation model for tropical cyclone tracks, with special emphasis on landfall behavior. *Nat. Hazards*, 73:335–353.

Kroese, D. P., Taimre, T., and Botev, Z. I. (2011). *Handbook of Monte Carlo Methods.* J. Wiley & Sons, Hoboken.

Krzysztofowicz, R. (1998). Point-to-area rescaling of probabilistic quantitative precipitation forecasts. *J. Appl. Meteorol.*, 38:786–796.

Lang, P. (2001). Cell tracking and warning indicators derived from operational radar products. In *Proc. 30th Int. Conf. Radar Meteorol.*, pages 207–211, Munich.

Lanza, L. G. (2000). A conditional simulation model of intermittent rain fields. *Hydrol. Earth Syst. Sci.*, 4:173–183.

Lawson, C. L. and Hanson, R. J. (1974). *Solving Least Squares Problems.* Prentice-Hall, Englewood Cliffs.

Ledo, A., Condés, S., and Montes, F. (2011). Intertype mark correlation function: A new tool for the analysis of species interactions. *Ecol. Model.*, 222:580–587.

Li, C. Q. (2000). A stochastic model of severe thunderstorms for transmission line design. *Probab. Eng. Mech*, 15:359–364.

Libby, W. F. (1955). *Radiocarbon Dating.* University of Chicago Press, Chicago.

Lück, S., Fichtl, A., Sailer, M., Joos, H., Brenner, R. E., Walther, P., and Schmidt, V. (2013). Statistical analysis of the intermediate filament network in cells of mesenchymal lineage by greyvalue-oriented segmentation methods. *Comput. Stat.*, 28:139–160.

Majewski, D. (1998). The new global icosahedral-hexagonal grid point model GME of the Deutscher Wetterdienst. In *ECMWF Semin. Proc.: Recent Dev. Numer. Methods Atmos. Model.*, pages 173–201, Reading.

Majewski, D., Liermann, D., Prohl, P., Ritter, B., Buchhold, M., Hanisch, T., Paul, G., Wergen, W., and Baumgardner, J. (2002). The global icosahedral-hexagonal grid point model GME: Description and high-resolution tests. *Mon. Weather Rev.*, 130:319–338.

Mayer, J., Schmidt, V., and Schweiggert, F. (2004). A unified simulation framework for spatial stochastic models. *Simul. Model. Pract. Theory*, 12:307–326.

McShea, W. and Healy, W. (2002). *Oak Forest Ecosystems.* John Hopkins University Press, Baltimore.

Meinhardt, M., Lück, S., Martin, P., Felka, T., Aicher, W., Rolauffs, B., and Schmidt, V. (2012). Modeling chondrocyte patterns by elliptical cluster processes. *J. Struct. Biol.*, 177:447–458.

Molchanov, I. S. (2005). *Theory of Random Sets.* Springer, London.

Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman & Hall/CRC, Boca Raton.

Montero, J.-M, Fernández-Avilés, G., and Mateu, J. (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging.* J. Wiley & Sons, Chichester.

Müller, W. G. (1999). Least-squares fitting from the variogram cloud. *Stat. Probab. Lett.*, 43:93–98.

Munoz, S. E. and Gajewski, K. (2010). Distinguishing prehistoric human influence on late-Holocene forests in southern Ontario, Canada. *Holocene*, 20:967–981.

Munoz, S. E., Gajewski, K., and Peros, M. C. (2010). Synchronous environmental and cultural change in the prehistory of the northeastern United States. *Proc. Natl. Acad. Sci.*, 107:22008–22013.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.*, 10:186–190.

Nadaraya, E. A. (1989). *Nonparametric Estimation of Probability Densities and Regression Curves*. Kluwer Academic Publishers, Dordrecht.

Neuhäuser, D., Hirsch, C., Gloaguen, C., and Schmidt, V. (2016). A stochastic model for multi-hierarchical networks. *Methodol. Comput. Appl. Probab.*, 18:1129–1151.

Neuman, S. P. and Jacobson, E. A. (1984). Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels. *Math. Geol.*, 16:223–240.

Neumann, M., Staněk, J., Pecho, O., Holzer, L., Beneš, V., and Schmidt, V. (2016). Stochastic 3D modeling of complex three-phase microstructures in SOFC-electrodes with completely connected phases. *Comput. Mater. Sci.*, 118:353–364.

Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *J. R. Stat. Soc. Ser. B*, 20:1–29.

Nguyen, H. T. (2006). *An Introduction to Random Sets*. Chapman & Hall/CRC, Boca Raton.

Northrop, P. J. (1998). A clustered spatial-temporal model of rainfall. *Proc. R. Soc. Lond. A*, 454:1875–1888.

O'Fallon, B. D. and Fehren-Schmitz, L. (2011). Native Americans experienced a strong population bottleneck coincident with European contact. *Proc. Natl. Acad. Sci.*, 108:20444–20448.

Onof, C. J., Chandler, R. E., Kakou, A., Northrop, P. J., Wheater, H. S., and Isham, V. S. (2000). Rainfall modelling using Poisson-cluster processes: A review of developments. *Stoch. Environ. Res. Risk Assess.*, 14:384–411.

Paciorek, C. J. and McLachlan, J. S. (2009). Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen proxy record. *J. Am. Stat. Assoc.*, 104:608–622.

Paschalis, A., Molnar, P., Fatichi, S., and Burlando, P. (2013). A stochastic model for high-resolution space-time precipitation simulation. *Water Resour. Res.*, 49:8400–8417.

Pedersen, M. W., Ruter, A., Schweger, C., Friebe, H., Staff, R. A., Kjeldsen, K. K., Mendoza, M. L. Z., Beaudoin, A. B., Zutter, C., Larsen, N. K., Potter, B. A., Nielsen, R., Rainville, R. A., Orlando, L., Meltzer, D. J., Kjær, K. H., and Willerslev, E. (2016). Postglacial viability and colonization in North America's ice-free corridor. *Nat.*, 537:45–49.

Peleg, N., Marra, F., Fatichi, S., Paschalis, A., Molnar, P., and Burlando, P. (2017). Spatial variability of extreme rainfall at radar subpixel scale. *J. Hydrol.* (in print).

Peros, M. C., Munoz, S. E., Gajewski, K., and Viau, A. E. (2010). Prehistoric demography of North America inferred from radiocarbon data. *J. Archaeol. Sci.*, 37:656–664.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

Quine, M. P. and Watson, D. F. (1984). Radial simulation of n-dimensional Poisson processes. *J. Appl. Probab.*, 21:548–557.

Ramesh, N. I., Thayakaran, R., and Onof, C. J. (2013). Multi-site doubly stochastic Poisson process models for fine-scale rainfall. *Stoch. Environ. Res. Risk Assess.*, 27:1383–1396.

Ramm, A. G. (2005). *Random Fields Estimation*. World Scientific, Singapore.

Reimer, P. J., Baillie, M. G. L., Bard, E., Bayliss, A., Beck, J. W., Bertrand, C. J. H., Blackwell, P. G., Buck, C. E., Burr, G. S., Cutler, K. B., Damon, P. E., Edwards, R. L., Fairbanks, R. G., Friedrich, M., Guilderson, T. P., Hogg, A. G., Hughen, K. A., Kromer, B., McCormac, G., Manning, S., Bronk Ramsey, C., Reimer, R. W., Remmele, S., Southon, J. R., Stuiver, M., Talamo, S., Taylor, F. W., van der Plicht, J., and Weyhenmeyer, C. E. (2004). IntCal04 terrestrial radiocarbon age calibration, 0–26 cal kyr BP. *Radiocarb.*, 46:1029–1058.

Reimer, P. J., Bard, E., Bayliss, A., Beck, J. W., Blackwell, P. G., Bronk Ramsey, C., Buck, C. E., Cheng, H., Edwards, R. L., Friedrich, M., Grootes, P. M., Guilderson, T. P., Haflidason, H., Hajdas, I., Hatte, C., Heaton, T. J., Hoffmann, D. L., Hogg, A. G., Hughen, K. A., Kaiser, K. F., Kromer, B., Manning, S. W., Niu, M., Reimer, R. W., Richards, D. A., Scott, E. M., Southon, J. R., Staff, R. A., Turney, C. S. M., and van der Plicht, J. (2013). Intcal13 and Marine13 radiocarbon age calibration curves 0-50,000 years cal BP. *Radiocarb.*, 55:1869–1887.

Rodriguez-Iturbe, I., Cox, D. R., and Eagleson, P. S. (1986). Spatial modelling of total storm rainfall. *Proc. R. Soc. Lond. A*, 403:27–50.

Rodriguez-Iturbe, I., Cox, D. R., and Isham, V. S. (1987). Some models for rainfall based on stochastic point processes. *Proc. R. Soc. Lond. A*, 410:269–288.

Rodriguez-Iturbe, I., Cox, D. R., and Isham, V. S. (1988). A point process model for rainfall: Further developments. *Proc. R. Soc. Lond. A*, 417:283–298.

Rumpf, J., Pickl, S., Elspaß, S., König, W., and Schmidt, V. (2010). Quantification and statistical analysis of structural similarities in dialectological area-class maps. *Dialectol. Geolinguist.*, 18:73–100.

Rumpf, J., Weindl, H., Höppe, P., Rauch, E., and Schmidt, V. (2009). Tropical cyclone hazard assessment using model-based track simulation. *Nat. Hazards*, 48:383–398.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Stat.*, 22:1346–1370.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC, Boca Raton.

Schilling, R. L. (2005). *Measures, Integrals and Martingales*. Cambridge University Press, Cambridge.

Schimek, M. G., editor (2000). *Smoothing and Regression*. J. Wiley & Sons, New York.

Schneider, R. and Weil, W. (2008). *Stochastic and Integral Geometry*. Springer, Berlin.

Scott, D. W. (1992). *Multivariate Density Estimation*. J. Wiley & Sons, New York.

Shimatani, K. (2002). Point processes for fine-scale spatial genetics and molecular ecology. *Biom. J.*, 44:325–352.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.

Skoglund, P. and Reich, D. (2016). A genomic view of the peopling of the Americas. *Curr. Opin. Genet. Dev.*, 41:27–35.

Smith, J. A. and Krajewski, W. F. (1987). Statistical modeling of space-time rainfall using radar and rain gage observations. *Water Resour. Res.*, 23:1893–1900.

Spettl, A., Werz, T., Krill III, C. E., and Schmidt, V. (2016). Stochastic modeling of individual grain behavior during Ostwald ripening at ultra-high volume fractions of the coarsening phase. *Comput. Mater. Sci.*, 124:290–303.

Srikanthan, R. and Pegram, G. G. S. (2009). A nested multisite daily rainfall stochastic generation model. *J. Hydrol.*, 371:142–153.

Steele, J. (2010). Radiocarbon dates as data: Quantitative strategies for estimating colonization front speeds and event densities. *J. Archaeol. Sci.*, 37:2017–2030.

Stenzel, O., Hirsch, C., Brereton, T. J., Baumeier, B., Andrienko, D., Kroese, D. P., and Schmidt, V. (2014). A general framework for consistent estimation of charge transport properties via random walks in random environments. *Multiscale Model. Simul.*, 12:1108–1134.

Stoyan, D. and Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields.* J. Wiley & Sons, Chichester.

Streit, R. L. (2010). *Poisson Point Processes: Imaging, Tracking and Sensing.* Springer, New York.

Surovell, T. A., Finley, J. B., Smith, G. M., Brantingham, P. J., and Kelly, R. (2009). Correcting temporal frequency distributions for taphonomic bias. *J. Archaeol. Sci.*, 36:1715–1724.

Takezawa, K. (2006). *Introduction to Nonparametric Regression.* J. Wiley & Sons, Hoboken.

Thayakaran, R. and Ramesh, N. I. (2017). Doubly stochastic Poisson pulse model for fine-scale rainfall. *Stoch. Environ. Res. Risk Assess.*, 31:705–724.

Thompson, R. S., Anderson, K. H., and Bartlein, P. J. (1999). *Atlas of Relations between Climatic Parameters and Distributions of Important Trees and Shrubs in North America.* US Geological Survey Professional Paper 1650A-B. USGS, Denver.

Turner, V., Gantz, J. F., Reinsel, D., and Minton, S. (2014). The digital universe of opportunities: Rich data and the increasing value of the internet of things. Technical report, IDC Analyze the Future.

Vale, T. R. (2002). The pre-European landscape of the United States: Pristine or humanized? In Vale, T. R., editor, *Fire, Native Peoples and the Natural Landscape*, pages 1–40. Island Press, Washington DC.

Viau, A. E., Gajewski, K., Sawada, M. C., and Fines, P. (2006). Millennial-scale temperature variations in North America during the Holocene. *J. Geophys. Res.*, 111:D09102.

Wackernagel, H. (2003). *Multivariate Geostatistics*, 3rd edition. Springer, Berlin.

Wand, M. P. (1992). Error analysis for general multivariate kernel estimators. *Nonparametr. Stat.*, 2:1–15.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman & Hall, London.

Wapler, K. (2017). The life-cycle of hail storms: Lightning, radar reflectivity and rotation characteristics. *Atmos. Res.*, 193:60–72.

Wapler, K., Goeber, M., and Trepte, S. (2012). Comparative verification of different nowcasting systems to support optimisation of thunderstorm warnings. *Adv. Sci. Res.*, 8:121–127.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya Ser. A*, 26:359–372.

Westhoff, D., Feinauer, J., Kuchler, K., Mitsch, T., Manke, I., Hein, S., Latz, A., and Schmidt, V. (2017). Parametric stochastic 3D model for the microstructure of anodes in lithium-ion power cells. *Comput. Mater. Sci.*, 126:453–467.

Wheater, H. S., Chandler, R. E., Onof, C. J., Isham, V. S., Bellone, E., Yang, C., Lekkas, D., Lourmas, G., and Segond, M.-L (2005). Spatial-temporal rainfall modelling for flood risk estimation. *Stoch. Environ. Res. Risk Assess.*, 19:403–416.

Wheater, H. S., Isham, V. S., Cox, D. R., Chandler, R. E., Kakou, A., Northrop, P. J., Oh, L., Onof, C. J., and Rodriguez-Iturbe, I. (2000). Spatial-temporal rainfall fields: Modelling and statistical aspects. *Hydrol. Earth Syst. Sci.*, 4:581–601.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*, 3rd edition. Academic Press, San Diego.

Wilks, D. S. and Eggleston, K. L. (1992). Estimating monthly and seasonal precipitation distributions using the 30- and 90-day outlooks. *J. Clim.*, 10:77–83.

Williams, J. W. and Shuman, B. N. (2008). Obtaining accurate and precise environmental reconstructions from the modern analog technique and North American surface pollen dataset. *Quat. Sci. Rev.*, 27:669–687.

Williams, J. W., Shuman, B. N., Webb III, T., Bartlein, P. J., and Leduc, P. L. (2004). Late-Quaternary vegetation dynamics in North America: Scaling from taxa to biomes. *Ecol. Monogr.*, 74:309–324.

Williams, M. (1989). *Americans and their Forests: A Historical Geography.* Cambridge University Press, New York.

Winterrath, T. and Rosenow, W. (2007). A new module for the tracking of radar-derived precipitation with model-derived winds. *Adv. Geosci.*, 10:77–83.

Winterrath, T., Rosenow, W., and Weigl, E. (2012). On the DWD quantitative precipitation analysis and nowcasting system for real-time application in German flood risk management. In *Proc. Int. Symp. Weather Radar Hydrol.*, pages 323–329, Exeter.

Woolhiser, D. A. and Osborn, H. B. (1985). A stochastic model of dimensionless thunderstorm rainfall. *Water Resour. Res.*, 21:511–522.

Yaglom, A. M. (1987a). *Correlation Theory of Stationary and Related Random Functions - 1. Basic Results.* Springer, New York.

Yaglom, A. M. (1987b). *Correlation Theory of Stationary and Related Random Functions - 2. Supplementary Notes and References.* Springer, New York.

Yang, C., Chandler, R. E., Isham, V. S., and Wheater, H. S. (2005). Spatial-temporal rainfall simulation using generalized linear models. *Water Resour. Res.*, 41:W11415.

Zängl, G., Reinert, D., Rípodas, P., and Baldauf, M. (2015). The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Q. J. R. Meteorol. Soc.*, 141:563–579.

Nomenclature

Latin symbols

$b(x, r)$	d-dimensional closed ball with center $x \in \mathbb{R}^d$ and radius $r > 0$, page 23
$\mathcal{B}$	σ-algebra of subsets of S, page 25
$\mathcal{B}^{\mathcal{T}}$	σ-algebra of subsets of $S^{\mathcal{T}}$, page 27
$\mathcal{B}(B)$	Borel σ-algebra generated by all open subsets of $B \in \mathcal{B}_0(\mathbb{R}^d)$, page 23
$\mathcal{B}(\mathbb{R}^d)$	Borel σ-algebra on $\mathbb{R}^d$, page 23
$\mathcal{B}_0(\mathbb{R}^d)$	family of bounded Borel sets in $\mathbb{R}^d$, page 23
C	family of closed sets in $\mathbb{R}^d$, page 58
$\mathcal{C}$	σ-algebra of subsets of C, page 58
cov	covariance, page 29
$\det(A)$	determinant of a matrix $A \in \mathbb{R}^{m \times m}$, page 24
d_{GC}	great-circle distance, page 33
$\mathrm{diag}(a_1, \ldots, a_m)$	diagonal matrix with entries $a_1, \ldots, a_m \in \mathbb{R}$ on the main diagonal, page 23
$\mathbb{E}$	expectation, page 29
F	family of Borel-measurable functions $f : \mathbb{R}^d \cup \{\infty\} \to [0, 1]$ such that $f(x) = 1$ for $x \notin B_f$ with some $B_f \in \mathcal{B}_0(\mathbb{R}^d)$, page 49
$\mathcal{F}$	σ-algebra of subsets of Ω, page 24
H	family of symmetric and positive definite matrices in $\mathbb{R}^{d \times d}$, page 63
$\mathcal{H}_f(x)$	Hessian matrix of a real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ with continuous first- and second-order partial derivatives at $x \in \mathbb{R}^d$, page 24

I	unit matrix, page 23
K	family of compact sets in $\mathbb{R}^d$, page 58
M	family of locally finite measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, page 46
$\mathcal{M}$	σ-algebra of subsets of M, page 46
$\mathbb{N}$	set of natural numbers, page 23
$\mathbb{N}_0$	set of nonnegative integers, page 23
N	family of locally finite and simple counting measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, page 47
$\mathcal{N}$	σ-algebra of subsets of N, page 47
o	origin in $\mathbb{R}^d$, page 23
$\mathbb{P}$	probability measure on $(\Omega, \mathcal{F})$, page 24
$\mathbb{R}$	set of real numbers, page 23
$\mathbb{R}^d$	set of d-dimensional vectors with real components, page 23
$\mathbb{R}^{m\times n}$	set of $m \times n$ matrices with real coefficients, page 23
S	state space of a random element X or a random function $\{X(t), t \in \mathcal{T}\}$, page 25
$S^{\mathcal{T}}$	set of functions with state space S and index set $\mathcal{T}$, page 27
$\mathbb{S}$	set of possible random errors in the forecast models of DWD, page 81
$\mathrm{supp}(f)$	support of a real-valued function $f : \mathbb{R}^d \to \mathbb{R}$, page 24
$\mathbb{T}$	set of thresholds $\{0, 0.1, 0.2, 0.3, 0.5, 0.7, 1, 2, 3, 5, 10, 15\}$ for precipitation amounts in mm, page 107
$\mathcal{T}$	index set of a random function $\{X(t), t \in \mathcal{T}\}$, page 26
$\mathrm{tr}(A)$	trace of a matrix $A \in \mathbb{R}^{m\times m}$, page 24
var	variance, page 29

Greek symbols

δB	rotation of $B \subset \mathbb{R}^d$ by $\delta : \mathbb{R}^d \to \mathbb{R}^d$, page 24
ν_d	d-dimensional Lebesgue measure, page 23
Ω	sample space, page 24
$\sigma(X)$	σ-algebra generated by a random element X, page 81
Θ	parameter space, page 44

Further symbols

A^{-1}	inverse of a matrix $A \in \mathbb{R}^{m\times m}$, page 23
$A^{\top}$	transpose of a matrix $A \in \mathbb{R}^{m\times n}$, page 23
$\#B$	cardinality of $B \subset \mathbb{R}^d$, page 24
$\overline{B}$	topological closure of $B \subset \mathbb{R}^d$, page 24
$X \overset{\mathrm{d}}{=} Y$	equality in distribution of random elements X and Y, page 25
$\lVert \cdot \rVert_d$	Euclidean norm on $\mathbb{R}^d$, page 23
$\mathbb{1}_B$	indicator function of $B \subset \mathbb{R}^d$, page 24
$\nabla_f(x)$	gradient of a real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ with continuous first- and second-order partial derivatives at $x \in \mathbb{R}^d$, page 24
$B + x$	translation of $B \subset \mathbb{R}^d$ by $x \in \mathbb{R}^d$, page 24
$B_1 \oplus B_2$	Minkowski sum of $B_1, B_2 \subset \mathbb{R}^d$, page 24

Abbreviations

BP	before present, page 166
BS	Brier score, page 98
BSS	Brier skill score, page 98
CARD	Canadian Archaeological Radiocarbon Database, page 166
DBSCAN	Density Based Spatial Clustering of Applications with Noise, page 148
DWD	Deutscher Wetterdienst, page 4
ECMWF	European Centre for Medium-Range Weather Forecasts, page 17
GLS	generalized least squares, page 44
GME	Globalmodell Europa, page 14
ICON	Icosahedral Nonhydrostatic, page 14
IFS	Integrated Forecasting System, page 17
KDE	kernel density estimator, page 64
KONRAD	Konvektionsentwicklung in Radarprodukten, page 13
LCV	likelihood cross-validation, page 73
LS	logarithmic score, page 140
LSS	logarithmic skill score, page 139
MISE	mean integrated squared error, page 65

MOS	Model Output Statistics, page 16
MSD	mean squared difference, page 127
NWE	Nadaraya-Watson estimator, page 70
NWP	numerical weather prediction, page 4
OLS	ordinary least squares, page 44
PDF	probability density function, page 60
PWP	probabilistic weather prediction, page 7
RADOLAN	Radar-Online-Aneichung, page 12
RPS	ranked probability score, page 133
RPSS	ranked probability skill score, page 132
STNSRP	spatio-temporal Neyman-Scott rectangular pulses, page 19
STREAP	space-time realizations of areal precipitation, page 20
US	United States, page 166
UTC	Universal Time, Coordinated, page 15

Abstract

In the present thesis, we develop novel stochastic models, statistical methods, and simulation algorithms for meteorological and paleogeographical space-time data. In this context, we aim to address two practical problems from different scientific disciplines. On the one hand, we consider an open question from meteorology, which has a high relevance in operational weather prediction. Currently applied probabilistic forecast methods are designed to estimate probabilities for the occurrence of weather events at fixed geographical locations, which is why they are called point probabilities in literature. However, for the issuing of weather warnings it often is desired to provide probabilities for the occurrence of certain weather events (e.g., heavy precipitation, thunderstorms) somewhere within an area, which are denoted as area probabilities. As no generally applicable formulas for the derivation of area probabilities based on point probabilities are available, we propose several spatial stochastic models that can be used for the operational computation of area probabilities in a general context. The second problem considered in this thesis concerns the statistical reconstruction of paleodemographic and paleoecological trends during the Holocene. In recent years, extensive databases of archaeological and paleoenvironmental samples have been compiled but only few attempts to estimate spatio-temporal distributions of, e.g., populations or vegetations on a continental scale were made. Therefore, another goal of this thesis is the development of suitable nonparametric methods to compute series of maps showing population densities and vegetation abundances for different plant taxa in North America over the past 13000 years. Furthermore, we also discuss an statistical analysis of relationships between estimated demographic and environmental developments.

The present thesis consists of three parts. In Part I, we provide some important preliminaries to motivate the contents of this thesis and to allow for a better understanding of the considered data and the applied mathematical models and methods. The introduction in Chapter 1 briefly describes the general context of the practical problems which we attempt to address using the statistical methodology proposed in Parts II and III. To further increase the comprehensibility of the stochastic models and statistical methods developed in Part II for the representation of several weather events, Chapter 2 provides an overview of systems, methods, and products applied in operation weather prediction. Additionally, this chapter gives a review on previous approaches to the computation of area probabilities and on existing stochastic models for precipitation cells and precipitation amounts. Chapter 3 contains a detailed description of the mathematical basics which are needed throughout this thesis. In particular, this includes an introduction of random fields with a focus on the characterization and estimation of dependency structures, a discussion of models from stochastic geometry, their properties, and simulation algorithms, and a review of nonparametric smoothing methods used in kernel density estimation and kernel regression.

Part II of this thesis describes the development of stochastic models for the representation of different weather events. In Chapter 4, we suggest to model precipitation cells using circular discs with random centers and a joint random radius. Furthermore, we propose suitable statistical methods to compute important model characteristics based on available point probabilities. This approach is extended in Chapter 5 by assigning a randomly scaled response function to each precipitation cell such that the summed response functions can be considered as a representation of precipitation amounts. Again, methods are developed to algorithmically compute model characteristics based on available point forecasts. A further generalization of the model for precipitation cells is described in Chapter 6, where a representation of thunderstorm cells based on cluster processes is provided. Besides forecasted point probabilities, this more complex model also requires thunderstorm records from past periods for model fitting but in return produces much more realistic realizations of thunderstorm events than previous approaches. All three models introduced in Part II can be used for the operational derivation of area probabilities in a general context. To evaluate forecast quality, obtained area probabilities are compared to high-resolution weather observations.

Finally, in Part III we discuss a spatio-temporal analysis of prehistoric population densities and vegetation abundances in North America. At first, Chapters 7 and 8 describe similar nonparametric approaches to the statistical estimation of population and vegetation intensities over the past 13000 years based on extensive databases of radiocarbon dates and fossil pollen samples. The proposed methodology produces spatially and temporally smoothed intensity maps every 100 years while accounting for several potential biases such as inhomogeneous sampling strategies, taphonomic loss, and boundary effects. We provide a brief discussion of obtained population intensity maps for most regions of North America and assess their significance by performing a sensitivity and robustness analysis. Vegetation intensity maps are estimated for a series of 10 plant taxa occurring in North America and an interpretation of results for one example taxon is given. In Chapter 9, we analyze statistical relationships between obtained population and vegetation intensities (as well as 500-year changes in both variables) by estimating spatio-temporal cross-correlation functions. A nonparametric method to assess the significance of correlation results is proposed and results for one example taxon are interpreted in the context of scientific literature.

The thesis is concluded by a summary of results and an outlook on open questions in Chapter 10 to motivate topics for future research.

Zusammenfassung

Die vorliegende Dissertation beschreibt die Entwicklung neuartiger stochastischer Modelle, statistischer Methoden und Simulationsalgorithmen für meteorologische und paläogeographische Raum-Zeit-Daten mit dem Ziel, Lösungen für zwei praktische Probleme aus unterschiedlichen wissenschaftlichen Fachrichtungen zur Verfügung zu stellen. Zum einen befasst sich diese Arbeit mit einer noch ungelösten meteorologischen Fragestellung, welche eine hohe Relevanz in der operationellen Wettervorhersage hat. Derzeit verwendete probabilistische Prognosemethoden liefern sogenannte Punktwahrscheinlichkeiten, d. h. sie schätzen Wahrscheinlichkeiten für das Auftreten von Wetterereignissen an festen geographischen Lokationen. Für die Bereitstellung von Wetterwarnungen ist es allerdings häufig von Interesse, Wahrscheinlichkeiten für das Auftreten bestimmter Wetterereignisse (z. B. Starkregen, Gewitter) irgendwo in einem bestimmten Gebiet zu bestimmen, welche als Flächenwahrscheinlichkeiten bezeichnet werden. Da keine allgemein anwendbaren Formeln zur Bestimmung von Flächenwahrscheinlichkeiten aus Punktwahrscheinlichkeiten bekannt sind, werden in dieser Arbeit diverse stochastische Modelle entwickelt, welche die operationelle Berechnung von Flächenwahrscheinlichkeiten unter verschiedenen Rahmenbedingungen ermöglichen. Die zweite betrachtete Problemstellung betrifft die statistische Rekonstruktion paläodemographischer und paläoökologischer Entwicklungen während des Holozäns. Obwohl in den vergangen Jahren umfangreiche archäologische und paläoökologische Datenbanken zusammengetragen wurden, existieren in der Literatur nur sehr wenige Ansätze zur Schätzung räumlich-zeitlicher Bevölkerungs- oder Vegetationsverteilungen auf einer kontinentalen Skala. Daher beschreibt diese Arbeit geeignete parameterfreie Methoden zur Schätzung von Populationsdichten und Vegetationsstrukturen in Nordamerika während der vergangenen 13000 Jahre. Darauf basierend wird schließlich eine statistische Korrelationsanalyse durchgeführt, um Zusammenhänge zwischen demographischen und ökologischen Entwicklungen zu identifizieren.

Die vorliegende Arbeit besteht aus drei Teilen. Teil I enthält einige wichtige Grundlagen, welche das Thema dieser Dissertation motivieren und ein besseres Verständnis der betrachteten Daten und verwendeten Verfahren ermöglichen sollen. Die Einleitung in Kapitel 1 beschreibt den Hintergrund der Fragestellungen, die mit Hilfe der in Teil II und III entwickelten Modelle und Methoden beantwortet werden sollen. Um die Verständlichkeit der Inhalte von Teil II noch weiter zu erhöhen, wird in Kapitel 2 eine kurze Einführung zu Systemen, Methoden und Produkten der operationellen

Wettervorhersage gegeben. Darüber hinaus bietet dieses Kapitel auch eine kurze Zusammenfassung vorheriger Ansätze zur Berechnung von Flächenwahrscheinlichkeiten und zur stochastischen Modellierung von Niederschlagszellen und Niederschlagsmengen. Kapitel 3 enthält schließlich eine detaillierte Darstellung der mathematischen Grundlagen, auf die im Verlauf dieser Arbeit wiederholt verwiesen wird. Dies umfasst eine Einführung in Zufallsfelder mit einem Schwerpunkt auf Charakterisierung und Schätzung von Abhängigkeitsstrukturen, eine Beschreibung verschiedener Modelle der stochastischen Geometrie (inklusive wichtiger Eigenschaften und Simulationsalgorithmen) und eine Übersicht zu Kerndichteschätzung und Kernel-Regression.

Teil II dieser Arbeit beschäftigt sich mit der Entwicklung stochastischer Modelle zur Darstellung verschiedener Wetterereignisse. Zunächst beschreibt Kapitel 4 einen Ansatz zur Modellierung von Niederschlagszellen als Kreisscheiben mit zufälligen Mittelpunkten und zufälligem Radius. Zusätzlich werden geeignete statistische Methoden vorgeschlagen, mit deren Hilfe wichtige Modellcharakteristiken basierend auf verfügbaren Punktwahrscheinlichkeiten bestimmt werden können. Dieser Ansatz wird in Kapitel 5 erweitert. Jeder Niederschlagszelle wird eine zufällig skalierte Kernfunktion zugewiesen, sodass die Summe der Kernfunktionen als ein Modell für Niederschlagsmengen interpretiert werden kann. Wie zuvor werden statistische Methoden zur Berechnung von Modellcharakteristiken mittels verfügbaren Punktvorhersagen entwickelt. Eine weitere Verallgemeinerung der Darstellung von Niederschlagszellen in Kapitel 4 wird in Kapitel 6 betrachtet, welches ein Modell für Gewitterzellen basierend auf räumlichen Cluster-Prozessen beschreibt. Neben prognostizierten Punktwahrscheinlichkeiten werden auch Gewitteraufzeichnungen vergangener Perioden zur Modellanpassung benötigt, dafür sind aber auch deutlich realistischere Realisierungen von Gewitterereignissen möglich als in vorherigen Ansätzen. Alle drei eingeführten Modelle können für die operationelle Schätzung von Flächenwahrscheinlichkeiten basierend auf wiederholter Monte-Carlo-Simulation verwendet werden. Um die Vorhersagequalität einzuschätzen, wird zudem ein Vergleich erhaltener Flächenwahrscheinlichkeiten mit flächendeckenden Wetterbeobachtungen durchgeführt.

In Teil III wird schließlich eine statistische Raum-Zeit-Analyse von prähistorischen Populationsdichten und Vegetationsstrukturen in Nordamerika durchgeführt. Zunächst beschreiben Kapitel 7 und 8 vergleichbare nichtparametrische Methoden zur Schätzung von Populations- und Vegetationsintensitäten der vergangenen 13000 Jahre. Basierend auf umfangreichen Radiokarbondaten und fossilen Pollenproben werden räumlich und zeitlich geglättete Intensitätskarten in Abständen von je 100 Jahren geschätzt, wobei verschiedene mögliche Fehlerquellen (inhomogene Beprobungsdichte, taphonomischer Verlust, Randeffekte) Berücksichtigung finden. Zusätzlich werden eine Interpretation der Populationsintensitätskarten und eine Sensitivitäts- und Robustheitsanalyse durchgeführt, um die Gültigkeit der erhaltenen paläodemographischen Ergebnisse zu untersuchen. Vegetationsintensitätskarten werden für 10 Pflanzentaxa geschätzt, wobei im Rahmen dieser Arbeit nur die Ergebnisse für ein Taxon genauer diskutiert werden. Zuletzt beschreibt Kapitel 9 eine statistische Korrelationsanalyse, in der räumlich-zeitliche Kreuzkorrelationsfunktionen zwischen erhaltenen Populations- und Vegetationsintensitäten (sowie zwischen 500-jährigen Veränderungen in beiden Variablen) untersucht werden. Es wird zudem eine nichtparametrische Methode entwickelt, um die Signifikanz erhaltener Korrelationen einschätzen zu können.

Zum Abschluss der Arbeit bietet Kapitel 10 eine Zusammenfassung der erzielten Ergebnisse und einen Ausblick auf offene Fragestellungen, um mögliche zukünftige Forschungsthemen zu motivieren.

Danksagung

An dieser Stelle möchte ich mich bei zahlreichen Personen bedanken, die durch ihre Unterstützung zum Gelingen der vorliegenden Dissertation beigetragen haben.

Zuerst gebührt mein Dank Prof. Dr. Volker Schmidt, der mich schon früh während meines Studiums förderte und mir die Möglichkeit geboten hat, an interessanten interdisziplinären Forschungsprojekten des Instituts für Stochastik mitzuwirken. Als Betreuer meiner Promotion hat er einen großen Anteil an den in dieser Arbeit beschriebenen Ergebnissen, nicht zuletzt durch seine vielen hilfreichen Ratschläge und Anmerkungen. Zudem hat mir Prof. Schmidt in den vergangenen Jahren diverse fachliche und soziale Kompetenzen vermittelt, wofür ich besonders dankbar bin.

Bei Prof. Dr. Viktor Beneš von der Karls-Universität in Prag möchte ich mich für sein Interesse in meiner Arbeit und die Bereitschaft zum Verfassen eines Gutachtens bedanken. Zudem danke ich ihm für seine Gastfreundschaft während meines Forschungsaufenthalts in Prag und die vielen Gespräche und Ratschläge, die zu den hier beschriebenen Resultaten beigetragen haben.

Des Weiteren möchte ich den vielen Projektpartnern meinen Dank aussprechen, mit denen ich in der Zeit meiner Promotion zusammenarbeiten durfte. Diese haben nicht nur großes Interesse für die theoretischen und praktischen Details meiner Arbeit gezeigt, sondern auch viel Zeit geopfert, um in zahlreichen Besprechungen Fragen zu Daten, Methoden und Hintergrundwissen aus den jeweiligen Anwendungsgebieten zu diskutieren und mir regelmäßig wertvolles Feedback zu geben.
Mein Dank gilt insbesondere Dr. Reinhold Hess vom Deutschen Wetterdienst für seine ausdauernde Unterstützung beim Initiieren und Bearbeiten der gemeinsamen Kooperationsprojekte sowie für das Bereitstellen vieler umfangreicher Datensätze. Zudem möchte ich mich bei Dr. Bernhard Reichert, Dr. Paul James, Dr. Andreas Veira, Dr. Jan Becker, Dr. Kathrin Wapler, Dr. Sebastian Trepte und Elmar Weigl vom Deutschen Wetterdienst bedanken, die mich mit ihrem Wissen und vielen Ratschlägen unterstützt haben.
Ein besonderer Dank gilt auch Prof. Dr. Konrad Gajewski, Prof. Dr. Rafal Kulik und Michelle Chaput von der Universität Ottawa für die vielen interessanten Diskussionen und ihre ansteckende Begeisterung für Themen der Archäologie und Paläoklimatologie. Außerdem möchte ich mich für ihre großartige Gastfreundschaft während meiner Besuche in Ottawa bedanken. Ich habe die Zeit in Kanada stets sehr genossen. Darüber

hinaus gilt mein Dank Dr. Matthew Betts vom Canadian Museum of History und Prof. Dr. Andrew Martindale von der University of British Columbia für ihr Interesse an meiner Arbeit.
Zuletzt möchte ich mich bei Dr. Zbyněk Pawlas und Antonín Koubek von der Karls-Universität in Prag für ihre Unterstützung und Ratschläge sowie für ihre Gastfreundschaft während meines Aufenthalts in Prag bedanken.

Ein herzliches Dankeschön möchte ich auch meinen (derzeitigen und ehemaligen) Kollegen am Institut für Stochastik aussprechen, die sowohl für fachliche Diskussionen als auch praktische Hilfe stets uneingeschränkt zur Verfügung standen. Ich habe das freundschaftliche Arbeitsklima am Institut und die vielen gemeinsamen Aktivitäten sehr genossen. Für die großartige Unterstützung und die vielen hilfreichen Ratschläge möchte ich mich in diesem Zusammenhang insbesondere bei Dr. Tim Brereton, Julian Feinauer, Dr. Gerd Gaiselmann, Lisa Handl, Dr. Christian Hirsch, Klaus Kuchler, Dr. David Neuhäuser, Matthias Neumann, Lukas Petrich, Dr. Ondřej Šedivý, Dr. Aaron Spettl, Dr. Ole Stenzel und Daniel Westhoff bedanken. Ich weiß eure Hilfe sehr zu schätzen. Außerdem gilt mein Dank Peter Schaumann und Raphael Wimmer für ihre Unterstützung bei diversen Programmierarbeiten und ihre hilfreichen Beiträge zu den von mir bearbeiteten Projekten.

Abschließend möchte ich meine große Dankbarkeit gegenüber meiner Familie für ihre liebevolle Unterstützung ausdrücken. Vielen Dank an meine Eltern, meine Großeltern und meinen Bruder, die sich stets für meine Arbeit interessiert haben und immer ein paar aufmunternde Worte parat hatten, wenn es etwas schleppender voran ging. Auch für ihre zahlreichen guten Ratschläge für alle Lebenslagen bin ich sehr dankbar. Ein besonders großes Dankeschön möchte ich meiner Frau Dominique aussprechen, die mich während meiner Promotion immer liebevoll unterstützt hat und jeden Tag für mich da ist. Sie hat immer an meine Fähigkeiten geglaubt, mich motiviert und mir auch geholfen mal von der Arbeit abzuschalten. Zuletzt möchte ich mich bei meiner wundervollen Tochter Amelia bedanken dafür, dass sie insbesondere nachts stets Rücksicht auf mich genommen hat und mich regelmäßig mit ihrem strahlenden Lächeln aufmuntert. Du bist die größte Motivation die man sich vorstellen kann!

Curriculum Vitae

Persönliche Daten

Name	Björn Kriesche
Anschrift	Wangener Straße 52, 89079 Ulm
E-Mail	`bjoern.kriesche@uni-ulm.de` `bjoernkriesche@hotmail.de`
Geboren	19.04.1988 in Stollberg/Erzgeb.
Familienstand	verheiratet, eine Tochter

Schulische und akademische Ausbildung

08/1994 – 07/2007	Besuch von Grundschule, Sekundarschule und Gymnasium in Hettstedt, Abschluss: Abitur
10/2007 – 09/2010	Studium der Wirtschaftsmathematik (B.Sc.) an der Universität Ulm, Thema der Abschlussarbeit: *Estimating the spatial correlations of a U.S. house price index*
10/2010 – 09/2012	Studium der Wirtschaftsmathematik (M.Sc.) an der Universität Ulm, Thema der Abschlussarbeit: *Improved stochastic modeling and simulation of tropical cyclone tracks at landfall*
seit 10/2012	Doktorand am Institut für Stochastik der Universität Ulm

Praktika und Berufserfahrungen

08/2009 – 10/2009	Praktikum bei der HDI-Gerling Leben in Köln, Abteilung Produkttechnik
12/2010 – 09/2012	Forschungsstudent am Institut für Stochastik der Universität Ulm
seit 10/2012	wissenschaftlicher Mitarbeiter am Institut für Stochastik der Universität Ulm

Forschungsaufenthalte

2013 – 2016	vier Forschungsaufenthalte an der Universität von Ottawa
10/2014	Forschungsaufenthalt an der Karls-Universität in Prag

Wissenschaftliche Publikationen

Artikel in referierten Fachzeitschriften

- B. Kriesche, H. Weindl, A. Smolka und V. Schmidt (2014): Stochastic simulation model for tropical cyclone tracks, with special emphasis on landfall behavior. *Natural Hazards* 73, 335-353.
- M. A. Chaput, B. Kriesche, M. Betts, A. Martindale, R. Kulik, V. Schmidt und K. Gajewski (2015): Spatiotemporal distribution of Holocene populations in North America. *Proceedings of the National Academy of Sciences of the United States of America* 112, 12127-12132.
- B. Kriesche, R. Hess, B. K. Reichert und V. Schmidt (2015): A probabilistic approach to the prediction of area weather events, applied to precipitation. *Spatial Statistics* 12, 15-30.
- A. Koubek, Z. Pawlas, T. Brereton, B. Kriesche und V. Schmidt (2016): Testing the random field model hypothesis for random marked closed sets. *Spatial Statistics* 16, 118-136.
- B. Kriesche, R. Hess und V. Schmidt (2017): A point process approach for spatial stochastic modeling of thunderstorm cells. *Probability and Mathematical Statistics* 37 (im Druck).
- B. Kriesche, A. Koubek, Z. Pawlas, V. Beneš, R. Hess und V. Schmidt (2017): On the computation of area probabilities based on a spatial stochastic model for precipitation cells and precipitation amounts. *Stochastic Environmental Research and Risk Assessment* 31, 2659-2674.

Artikel in Konferenzbänden

- B. Kriesche, A. Koubek, Z. Pawlas, V. Beneš, R. Hess und V. Schmidt (2015): A model-based approach to the computation of area probabilities for precipitation exceeding a certain threshold. *Proceedings of the 21st International Congress on Modelling and Simulation*, Gold Coast, Australien, 2103-2109.

Eingereichte Artikel

- K. Gajewski, B. Kriesche, M. A. Chaput, R. Kulik und V. Schmidt: Human-vegetation interactions during the Holocene in North America. *Ecological Monographs* (eingereicht).
- R. Hess, B. Kriesche, P. Schaumann, B. K. Reichert und V. Schmidt: Area precipitation probabilities derived from point forecasts for operational weather and warning service applications. *Quaterly Journal of the Royal Meteorological Society* (eingereicht).
- B. Kriesche, M. A. Chaput, R. Kulik, K. Gajewski und V. Schmidt: Estimation of spatio-temporal correlations of prehistoric population and vegetation in North America. *Geographical Analysis* (eingereicht).

Vorträge und Posterpräsentationen

Vorträge

- Modellbasierte Prognose gebietsbezogener Niederschlagswahrscheinlichkeiten. Projekttreffen beim DWD, 06.02.2013, Offenbach, Deutschland.

- Stochastic simulation model for tropical cyclone tracks: improvements of landfall behavior. 4th International Summit on Hurricanes and Climate Change, 16.06.2013, Kos, Griechenland.

- A probabilistic approach to the prediction of area-related weather events. 11th German Probability and Statistics Days, 04.03.2014, Ulm, Deutschland.

- Stochastic modeling of spatially resolved data, with applications to the prediction of area weather events. Seminar an der Universität von Ottawa, 18.09.2014, Ottawa, Kanada.

- Stochastic modeling of spatially resolved data, with applications to the prediction of area weather events. Seminar an der Karls-Universität in Prag, 21.10.2014, Prag, Tschechien.

- On the computation of area precipitation exceedance probabilities by modeling precipitation amounts. STOCHASTIKA 2015, 10.02.2015, Kohutka, Tschechien.

- A model-based approach to the computation of area probabilities for precipitation exceeding a certain threshold. 21st International Congress on Modelling and Simulation, 01.12.2015, Gold Coast, Australien.

- A model-based approach to the computation of area probabilities for precipitation exceeding a certain threshold. 13th International Meeting on Statistical Climatology, 08.06.2016, Canmore, Kanada.

- Stochastische Geometrie in der probabilistischen Wettervorhersage: Nutzung von kombinierten Daten aus MOSMIX und NowCastMIX. EMF-Statustreffen beim DWD, 14.07.2016, Offenbach, Deutschland.

- Stochastische Geometrie in der probabilistischen Wettervorhersage: Kombinierte Nutzung von Daten aus ModelMIX und NowCastMIX zur Prognose von Gewitterereignissen im Kürzestfristbereich. EMF-Statustreffen beim DWD, 10.05.2017, Offenbach, Deutschland.

Poster

- A model-based approach to the computation of area probabilities for precipitation exceeding a certain threshold. Workshop on Uncertainty Modeling in the Analysis of Weather, Climate and Hydrological Extremes, 16.06.2016, Banff, Kanada.

Erklärung

Hiermit versichere ich, Björn Kriesche, dass ich die vorliegende Arbeit selbstständig und nur mit den angegebenen Hilfsmitteln angefertigt habe. Alle Stellen, die wörtlich oder inhaltlich anderen Arbeiten entnommen sind, wurden durch Angabe der Quellen kenntlich gemacht. Die Satzung der Universität Ulm zur Sicherung guter wissenschaftlicher Praxis wurde beachtet. Ich erkläre außerdem, dass die von mir vorgelegte Dissertation bisher nicht im In- oder Ausland in dieser oder ähnlicher Form für ein anderes Promotionsverfahren eingereicht wurde. Ich versichere ferner die Richtigkeit der im Lebenslauf gemachten Angaben.

Ulm, den 24.10.2017

(Unterschrift)