exPrior: An R Package for the Formulation of Ex-Situ Priors

by Falk Heße, Karina Cucchi, Nura Kawa, and Yoram Rubin

Abstract The **exPrior** package implements a procedure for formulating informative priors of geo-statistical properties for a target field site, called *ex-situ priors* and introduced in Cucchi et al. (2019). The procedure uses a Bayesian hierarchical model to assimilate multiple types of data coming from multiple sites considered as similar to the target site. This prior summarizes the information contained in the data in the form of a probability density function that can be used to better inform further geostatistical investigations at the site. The formulation of the prior uses ex-situ data, where the data set can either be gathered by the user or come in the form of a structured database. The package is designed to be flexible in that regard. For illustration purposes and for easiness of use, the package is ready to be used with the worldwide hydrogeological parameter database (WWHYPDA) Comunian and Renard (2009).

Introduction

Characterizing the subsurface of our planet is an important task in fields such as geology, hydrogeology, and soil sciences. Yet compared to many other fields, the characterization of the subsurface is always burdened by large uncertainties. These uncertainties are caused by the general lack of data and the large spatial variability of many subsurface properties. The need to represent this uncertainty has led to the development of the field of geostatistics, wherein parameter values are treated as random variables defined by their probability distribution function (PDF). Today, the field of geostatistics has reached a mature state with many textbooks on the topic (Pyrcz and Deutsch, 2002; Rubin, 2003; Kitanidis, 2008) and a solid number of software tools being available for practitioners (for an overview, see, e.g. Rubin et al. (2018)). For the R language (R Core Team, 2014), a solid ecosystem for geostatistical analysis has evolved in the last years (Slater et al., 2019). Packages like **geoR** (Ribeiro and Diggle, 2001), **gstat** (Pebesma, 2004), **georob** (Papritz et al., 2014), and **RGeostats** (MINES ParisTech / ARMINES, 2019) provide a large collection of tools for geostatistical analysis. Moreover, geostatistical databases can be conveniently accessed with packages like **aqp** (Beaudette et al., 2013) and textbooks on geostatistics are starting to provide all their examples in R code (Diggle and Ribeiro, 2007; Banerjee et al., 2014).

Bayesian statistics provides the most appropriate framework to characterize uncertainty in general (Heße et al., 2019a). Bayesian methods are able to combine and assimilate data from disparate sources and jointly represent the different forms of uncertainty. As a result, Bayesian methods are nowadays increasingly employed in geostatistics and software implementations come as either standalone versions (Vrugt et al., 2009; Rubin et al., 2010) or R packages like **spBayes** (Finley et al., 2015), **R-INLA** (Lindgren and Rue, 2015), **spTimer** (Bakar and Sahu, 2015), **BayesNSGP** (Risser and Turek, 2020), and **anchoredDistr** (Savoy et al., 2017).

Yet, there is no package to date, which would provide such tools with the necessary foundation, i.e. prior distributions for the modeled quantities. Since the prior is the first step of any Bayesian analysis, its overall importance can hardly be overstated. Moreover, the ability of prior distributions to represent available background information in a given field makes them an important source of information that should not be neglected. Integrating them into a Bayesian workflow should be straightforward since most packages for Bayesian inference allow users to specify their prior distributions. In addition, the use of informative prior distributions in this field is easy to motivate. First, the parameters of geostatistical models are typically not simple convenience parameters but are part of physically-based partial differential equations. As a result, they correspond to real-world, physical measurements, making it possible to calibrate their prior distributions against empirical frequencies. Second, geostatistical models are usually site specific, making it conceptually easy to discriminate between the case-specific data, which should be used to compute the likelihood, and background data, which could be used to compute the prior distribution. In geostatistics, they are often called in-situ and ex-situ data, respectively. Calibrating the prior against ex-situ data only, therefore guarantees a clear separation between likelihood and prior.

To provide practitioners therefore with a tool for prior derivation, this paper introduces the R package **exPrior** (Heße et al., 2019b). It implements the derivation of ex-situ priors, i.e., statistical distributions of subsurface properties at a given site from ex-situ data collected at similar sites, following the Bayesian hierarchical model developed by Cucchi et al. (2019). The implementation is based on the **nimble** package (de Valpine et al., 2017) itself based on the BUGS language (Lunn et al., 2009). The objective of the **exPrior** package is to provide a ready-to-use software tool for

assimilating ex-situ data into ex-situ priors. This will encourage the use of informative distributions for practitioners in geosciences that might not be experts in Bayesian hierarchical models and may find it therefore difficult to work with them otherwise. The focus of this package is on the Gaussian process (GP) modelling paradigm. Although criticism exists, it is by far the most widely used paradigm, and a wide range range of software tools exist for the modeling of GPs (Pebesma, 2004). Functions in **exPrior** provide wrappers around **nimble** functions implementing the Bayesian data assimilation framework. Non-expert practitioners can therefore apply this method without needing to implement the model itself. Moreover, the package is tightly integrated with two other R packages that help to expand its functionality. First, the **geostatDB** package (Heße et al., 2019c) provides access to a large data set from the worldwide hydrogeological parameter database (WWHYPDA) Comunian and Renard (2009). Second, the **siteSimilarity** package (Kawa et al., 2020) allows for clustering of similar sites and therefore facilitates a further reduction of uncertainty.

Due to the background of the authors, the examples and data are drawn from stochastic hydrogeology, i.e., the field of geostatistics concerned with the statistical characterization of groundwater systems. Yet, the presented package is not confined to this field, and simply using other data or making slight revisions to the hierarchical model will quickly make the workflow amendable to other fields of geostatistics as well.

To familiarize the reader with the package, we start in the following by explaining the workflow for formulating informative prior distributions, where the prior distribution at an unexplored site is based on data collected from other sites. After that, we explain the package by detailing its structure and functionalities. Finally, we present several examples of prior derivation based on synthetic data and on an established database for hydrogeological parameters (Comunian and Renard, 2009).

Ex-Situ Priors

Let us assume that we want to model a specific geostatistical variable x at a target site S_0. Examples would be hydraulic conductivity, porosity, or permeability. To account for the unavoidable uncertainty, this variable should be modeled as a random variable X (Pyrcz and Deutsch, 2002; Rubin, 2003; Kitanidis, 2008). The simplest way to characterize this variable statistically is through its distribution $p(x)$. Yet, this would leave out any spatial correlations, so most geostatistical analyses try to account for them by using spatial random field models, typically a GP (Rasmussen and Williams, 2006; Gelfand and Schliep, 2016). Since such models are fully defined by their parameter vector θ, the aim of Bayesian inference is to use available, in-situ data y_{in} and derive the posterior distribution over these parameters $p(\theta|y_{in})$. This posterior represents a compromise between likelihood $p(y_{in}|\theta)$ and the prior distribution $p(\theta)$, with the likelihood representing the impact of the in-situ data. This, however, leaves open the specification of the prior distribution.

By definition, the prior distribution characterizes the knowledge about target parameters before observing in-situ data y_{in}. Therefore, y_{in} cannot be used for the definition of the prior (Berger, 2006). On the other hand, using no data and making the prior distribution as vague as possible seems far too prudent since this would ignore the wealth of background knowledge which exists for virtually any geostatistical variable. Such background knowledge can come from data collected at other sites $S_i, i \in 1 \ldots I$ (see schematic in Figure 1). To distinguish them from the site-specific, in-situ data y_{in}, we use the term ex-situ data y_{ex}. Our prior pdf for the parameters at a new site S_0 could therefore be based on these ex-situ data $p(\theta|y_{ex})$. To determine this $p(\theta|y_{ex})$, we propose the use of a dedicated statistical model (more on this below). By virtue of this model, the transfer of information from known donor sites S_i to a new site S_0 is a case of Bayesian prediction

$$p(\theta|y_{ex}) = \int_{\vartheta} p(\theta|\vartheta) p(\vartheta|y_{ex}) d\vartheta. \tag{1}$$

According to Eq. 1, the prior distribution $p(\theta|y_{ex})$ for a new site S_0 is the posterior predictive distribution of all S_i w.r.t. S_0 (see schematic in Figure 1). Mathematically, this means $p(\theta|y_{ex})$ is derived by weighing each single predictive distribution $p(\theta|\vartheta)$ with its corresponding posterior distribution $p(\vartheta|y_{ex})$ and marginalizing over the parameters ϑ. Please note the difference between the model for X at site S_0, say a GP defined by θ, and the model used for the transfer of data between sites defined by ϑ. Since geostatistical data are hierarchical in nature, this model should be hierarchical too (Kruschke, 2010; Gelfand, 2012; Gelman et al., 2013).

Formulation of the hierarchical model

In geostatistics, a common way to conceptualize a hierarchical ordering of the data is by using two levels (see schematic in Figure 1). The first level represents the population of the (spatially distributed)
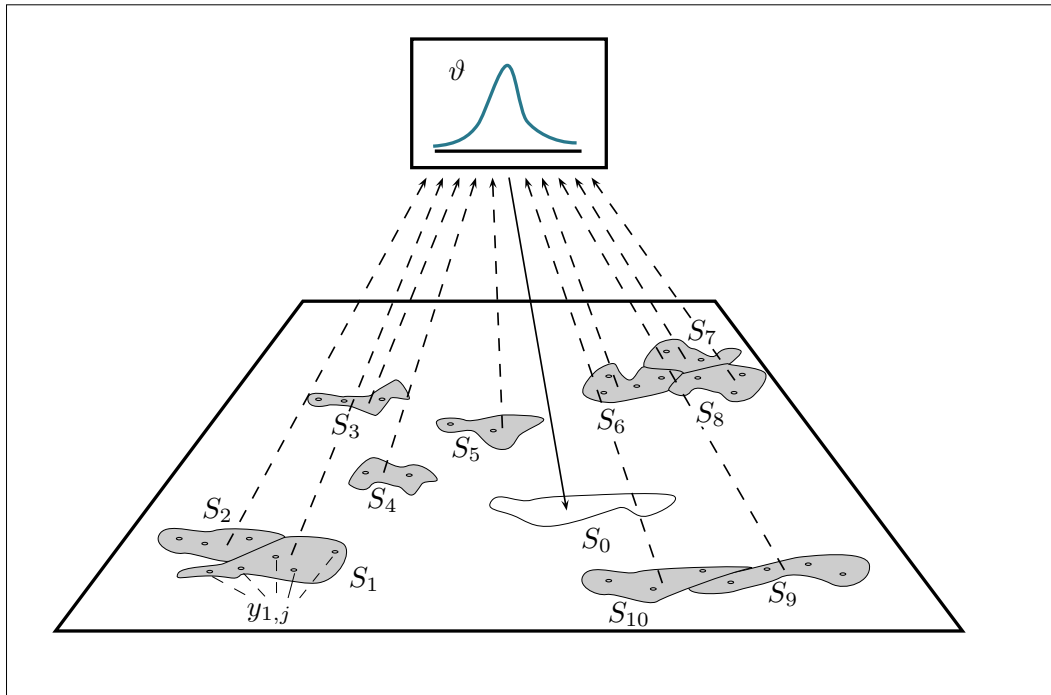
Figure 1: Schematic of the transfer of data from a number of donor sites S_i to a new site S_0. Measurement locations are denoted by circles. The statistical model used for the transfer is denoted by its parameter vector ϑ.

local data $y_{i,j}$ available on every given site S_i whereas the second level represents the global population of all available sites. In such a two-level hierarchical model, this relationship is represented such that the set of parameters ϑ is split up into two subsets $\vartheta = (\phi, \eta)$ for level 1 and 2 respectively. The hierarchical relationship between these two levels is represented by factorizing their joint probability through the definition of conditional probability $p(\phi, \eta) = p(\phi|\eta)p(\eta)$. The general formulation of the two-level model would then look like the following:

$$y_{i,j} \sim p(y|\phi_i), \tag{2a}$$

$$\phi_i \sim p(\phi|\eta), \tag{2b}$$

$$\eta \sim p(\eta). \tag{2c}$$

This means that each datum $y_{i,j}$ is drawn from a distribution with parameters ϕ_i, which are specific to site S_i only. This distribution, therefore, represents the local variability of the data found within a given site or intra-site variability (Eq. 2a). These local parameters ϕ_i are, in turn, drawn from a distribution specified by the global parameters called *hyperparameters* η. These hyperparameters, therefore, represent the global variability between sites or inter-site variability (Eq. 2b).

This general formulation allows to flexibly choose parametric models used for all distributions. Since $p(y|\phi_i)$ represents the data, this distribution should fit the empirically observed frequencies of y. Depending on the geostatistical parameter of interest, a user can use, e.g., the normal, log-normal, multivariate normal, or truncated normal distributions to model parameter behavior. Choosing the distributions of the hierarchical model itself, i.e., $p(\phi|\eta)$ and $p(\eta)$ is less straightforward and should reflect of mixture of the domain knowledge and statistical expertise. To exemplify this procedure, let us look at measurements of hydraulic conductivity. These data are often modeled with a log-normal distribution (Hoeksema and Kitanidis, 1985), while $p(\phi|\eta)$ can be modeled as a normal distribution (Gelman et al., 2013). The parametric form of $p(\eta)$ should be specified as vague priors initially, with the posteriors being determined by the data y_{ex}. Transforming our data into their log-normal form, as often done, the resulting hierarchical model would then be

$$y_{i,j} \sim \mathcal{N}(\mu_i, \sigma^2), \tag{3a}$$

$$\mu_i \sim \mathcal{N}(\alpha, \tau), \tag{3b}$$

$$(\sigma^2, \alpha, \tau) \sim p(\sigma^2)p(\alpha)p(\tau). \tag{3c}$$

In this example, we assumed that the variance σ^2 at each site is the same, making the local and global parameters identical. This assumption is, of course, a simplification but allows to reduce the number of parameters to be inferred.

Generation of the ex-situ prior distribution

Once the target variable is specified, and the ex-situ data y_{ex} collected, three steps are necessary to actually calculate the prior distribution. First, the user has to decide on the parametric model for the distributions and therefore fully specify the hierarchical model according to Eq. 2. Next, the posterior distributions of the parameters $p(\phi, \eta | y_{ex})$ are inferred. In **exPrior**, this is done using the Markov chain Monte Carlo (MCMC) implementation of **NIMBLE**. Finally, the ex-situ prior can be determined as the posterior predictive distribution (see Equation 1).

To familiarize ourselves with this procedure, let us look at these three steps in more detail.

Specifying a hierarchical model The specification of the hierarchical model in **exPrior** is done in BUGS code wrapped within the **NIMBLE** function `nimbleCode()`. The results are R objects from the BUGS models. In **NIMBLE**, every model is represented as a Directed Acyclic Graph (DAG), where each declaration in the model is a node which can be either deterministic or stochastic. Nodes are represented as vertices of a DAG, with edges connecting nodes implying dependence relationships.
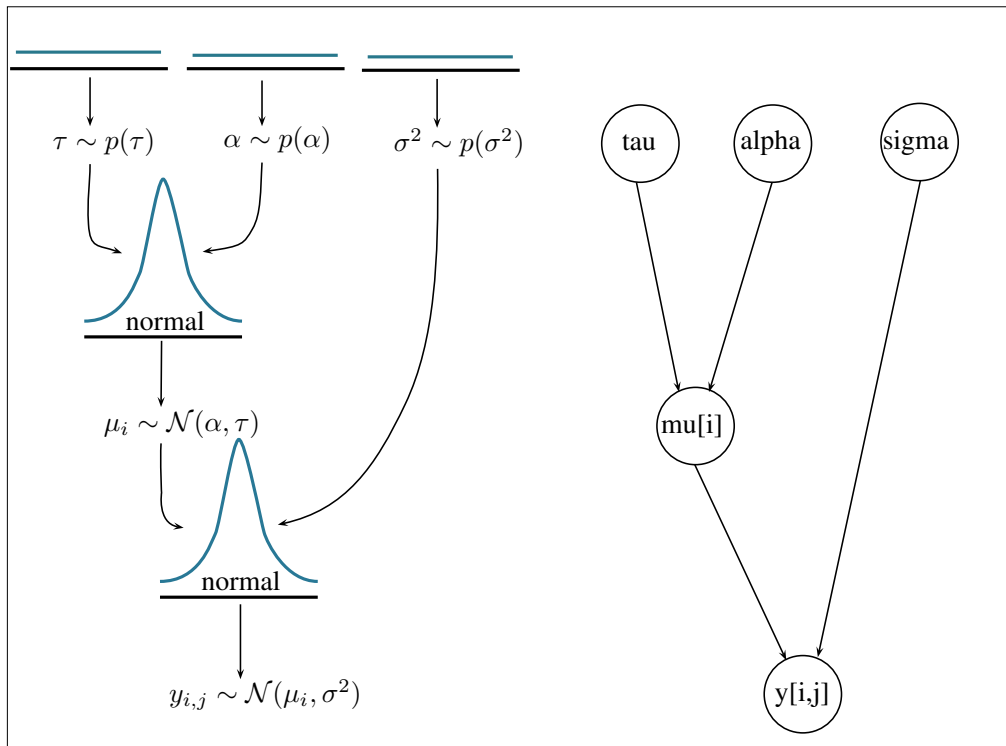

Figure 2: Schematic of the example model showing the statistical model on the left as defined in see Eq. 3 and the corresponding DAG on the right. The arrows show the hierarchical relationships between the variables. Both the ex-situ data $y_{i,j}$, and the site-specific mean μ_i are drawn from normal distributions. The hyperparameters $\eta = (\tau, \alpha, \sigma^2)$ are given initially vague hyperpriors to be updated later by the ex-situ data.

To exemplify this, let us consider the aforementioned model for the log-hydraulic conductivity. As mentioned, the data at each site are modeled as being drawn from a normal distribution with every site having the same variance σ^2 but site-specific means. This mean value is again drawn from a normal distribution. Accordingly, α is the global mean of the local means, and τ^2 represents the

global variability of the local means. Since no data are used at this point, the hyperpriors of the hyperparameters $\eta = (\alpha, \tau, \sigma)$ should be non-informative.

The model is therefore declared with five variables: alpha, tau, sigma, mu, and y. Once compiled, the model contains multiple nodes: one node for each hyperparameter alpha, tau, and sigma, I nodes for the site-specific means mu[i], and $\sum_i J_i$ nodes for the site measurements y[i,j]. In this model, the dependent nodes are the ex-situ data $y_{i,j}$, the μ_i values are estimated from these deterministic $y_{i,j}$. Similarly, the α and τ are estimated from the μ_i. In cases where the data are provided to genExPrior() in the form of moments (more on this later), mu[i] is deterministic, as the site-specific means are already provided in numeric form and not considered to be realizations from any distribution. The following pseudocode illustrates how the model is declared:

```
## declaration of a hierarchical model in nimbleCode

# 1. declare prior distributions for hyperparameters
hyperparameters ~ disn(...)

# 2. declare distributions for site-specific means
for i in 1:I {
  mu[i] ~ disn(hyperparam1, hyperparam2, ...) # distribution of mean mu at site i

# 3. declare distributions for measurements
  for j in 1:J {
  y[i,j] ~ disn(mu[i], hyperparam3, ...)
  }
}
```

This pseudocode model denotes the mathematical formulation of the hierarchical model in BUGS code, where the parameter for site-specific means μ_i is mu[i], and site-specific variances σ_i^2 is sigma2[i]. First, hyperparameters alpha and tau are assigned non-informative hyperpriors. Next, the code loops through each site and assigns the mean mu[i] a distribution whose parameters are hyperparameters. Finally, each of the J measurements y[i,j] at each site i is assigned a distribution with parameters mu[i] and hyperparameters.

The flexibility of this formulation allows the data y_{ex} to be assimilated in a full MCMC hierarchical model, as explained in the next section.

Estimating posterior values of parameters in the hierarchical model   Parameters in the hierarchical model are estimated from the data provided by the user, using MCMC. Once the model and the variables are declared, the hierarchical model is compiled using nimble::compileNimble(). The MCMC object is configured, built, and compiled using nimble::configureMCMC(), nimble::buildMCMC(), and nimble::compileNimble(), respectively and run using the run method of the compiled object. This method calculates an MCMC chain, the result of the estimation step. To improve numerical efficiency, NIMBLE includes a library of algorithms and a compiler that generates C++ for declared models and functions. Once a model is declared, nimbleCode is generated as C++ code, compiled, and reloaded into R.

The values in the compiled model are declared with data that are supplied by the user when running the function. For example, if a user inputs a data frame of measurements, each of the y[i,j] is defined with its corresponding data point. If a user provides moments, then each of the mu[i] is defined with its provided value. Once a model is compiled, genExPrior() envokes the MCMC sampler, which estimates the posterior distributions of the parameters.

Predicting the prior   Now that the ex-situ data are assimilated, our hierarchical model is fully conditioned on all available data. Normally, this would conclude a Bayesian inference. Yet as explained above, the posterior distribution has to be used to compute the predictive posterior distribution of the target variable at a new site (see Eq 1). In **exPrior**, this is simply done by drawing realizations of the target variable from distributions specified by the hierarchical model and parameterized by values sampled from the MCMC chains from the previous step. The final predictive posterior distribution is then estimated using kernel density estimation.

Associated packages

To support the functionality of **exPrior**, we provide two additional R packages on the GitHub account. First, the **geostatDB** package provides real-world data on subsurface measurements, which can

therefore be used directly for the derivation of informed prior distributions. Second, the **siteSimilarity** package allows users to determine a cluster of similar sites to focus on the most relevant data only and therefore reducing the overall uncertainty. Since both these packages provide benefits independently of **exPrior**, they will remain independent for the foreseeable future.

Real-world data: The geostatDB package

In this package, we provide real-world, geostatistical data from the WorldWide HYdrogeological Parameters DAtabase *WWHYPDA* (Comunian and Renard, 2009). This database has been designed to store values of the most important properties of earth materials and has been developed with the purpose of offering a complement of information in hydrogeological studies where there is a lack of data. To the best of the authors' knowledge, the WWHYPDA is the largest open-source database of hydrogeological parameters. Currently it contains a total of 20,523 subsurface measurements of 6 geostatistical parameters spanning 128 sites. A complete description and schematic of the database in its original form can be found in Comunian and Renard (2009).

Due to its size, as well as to facilitate further additions to the database, we use a dedicated R package **geostatDB** to host the data from WWHYPDA. This package is not yet on CRAN, but the latest release version can be found on https://github.com/GeoStat-Bayesian/geostatDB (Heße et al., 2019c). To exemplify the usage and impact of real-world data, we furnished **exPrior** with an example data set on porosity values from sandy aquifers. Its usage is described in Section 2.4.2 below.

geostatDB itself includes the WWHYPDA as an SQLite database, converted from the online MySQL database. The reasons for using SQLite are efficiency and accessibility. First, both SQL and SQLite databases can be easily read into R, while maintaining their original structure. SQLite specifically can be included in R packages without the need for a server, making it accessible to the user. A downside of this decision is that a user who updates the SQLite database in **geostatDB** does so without making changes to the original database, hosted on https://www.wwhypda.org. Thus, those who wish to contribute to the WWHYPDA are currently encouraged to do so by submitting data online.

It is important to note that data quality steps need to be implemented before applying the statistical algorithm to this database. Figure 3 contains two visualizations that describe the data present in the WWHYPDA, created in R. This visualization was done using the function getData() from the **geostatDB** package. The code can be found in the associated vignette of the package at https://github.com/GeoStat-Bayesian/geostatDB/blob/master/vignettes/explore_data.Rmd.

The notion of site similarity: The siteSimilarity package

In order to reduce the uncertainty in the prior distribution as much as possible, it is beneficial to focus only on data coming from sites which are similar to the one under investigation. It is therefore crucial to have a sound notion of site similarity. The **siteSimilarity** package uses hierarchical agglomerative clustering to categorize sites into clusters based on observable characteristics, such as environment type or rock type. Using the schematic in Figure 1, only those sites similar to site S_0 would be used as donor sites. Currently, the clustering achieves only a modest reduction in uncertainty when using a leave-one-out validation. This is caused by the overall limited number of sites, which, after clustering, get even more reduced. Yet the algorithm already provides the user with a tangible benefit, which is projected to increase as more and larger data sets with more sites become available.

Examples

Having now formally explained the workflow and associated packages of **exPrior**, we will illustrate said workflow with a series of examples. Starting with an easy inference problem, we will then explain how to use the included data, how to account for autocorrelation, and finally how to use soft data, in particular bounds, for the inference.

Please note that all examples given in the following only refer to the distribution of the expected quantity itself, e.g., porosity. This does, however, not mean that the parameters of a GP cannot be inferred since μ and σ are nodes in the hierarchical model and can be derived from it. On the other hand, higher-order statistics, like correlation length or anisotropy, are usually considered homogeneous across a given site and are consequentially not hierarchical. They can, therefore, be inferred using the classical estimation procedure.

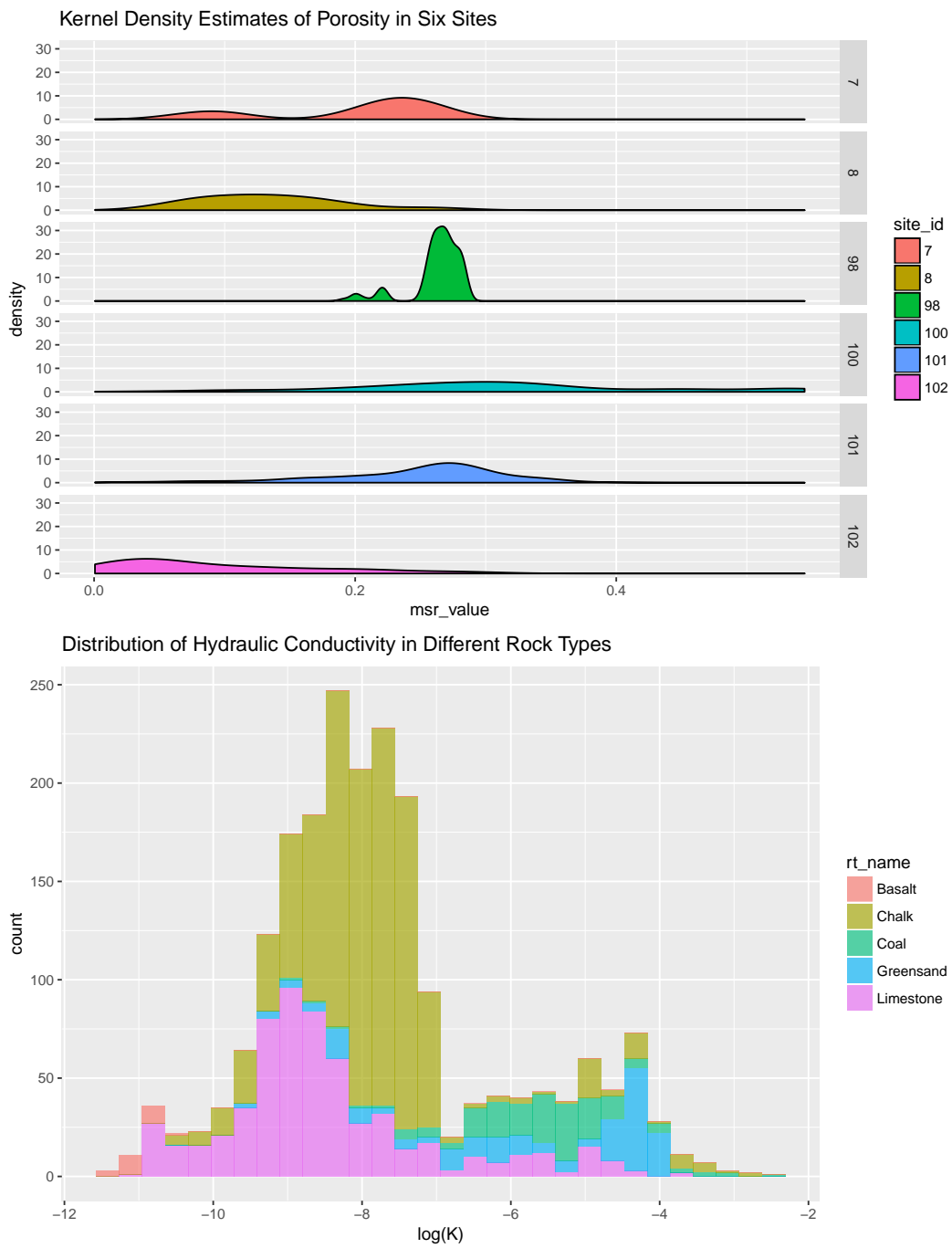The following four examples correspond to four vignettes, which can be found on the GitHub account of the **exPrior** package at https://github.com/GeoStat-Bayesian/exPrior/blob/master/

Figure 3: The two visualizations are made using the data from the WWHYPDA, obtained using the function getData() from the **geostatDB** package. The first figure shows the distribution of porosity values at several sites, derived using kernel density estimation. The second figure shows a set of histograms describing the distribution of hydraulic conductivity values for different material types.
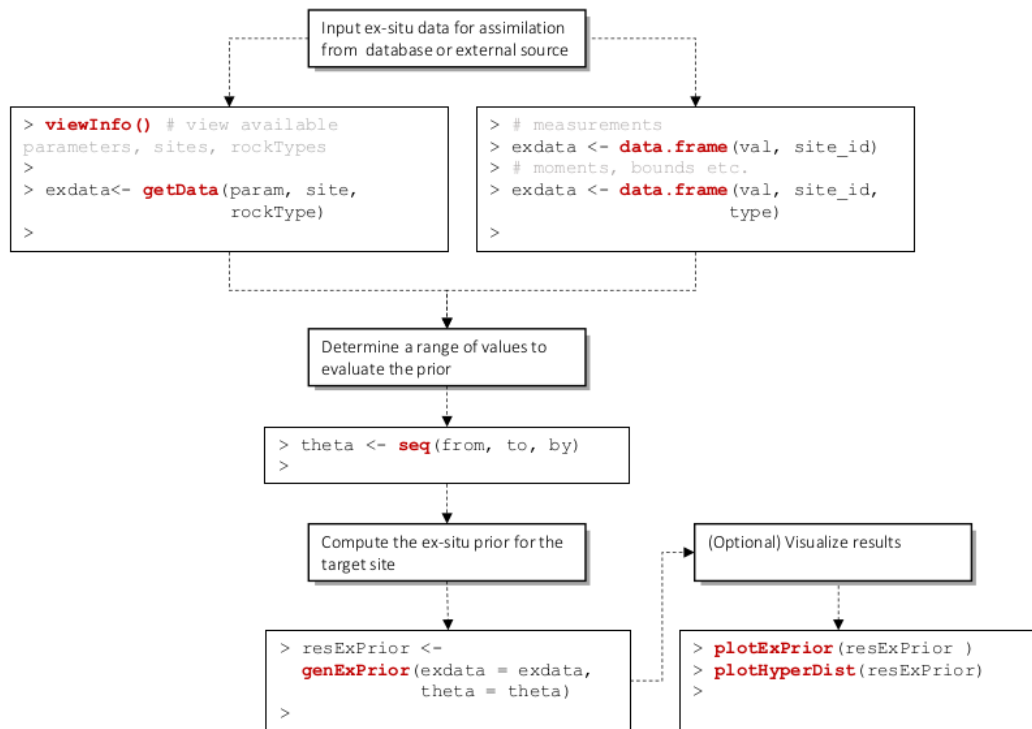
`vignettes`.

Figure 4: A flowchart showing the workflow of the user of **exPrior**. A user first expresses ex-situ data as an R dataframe. The user then determines the range of values at which to compute a prior (usually the minimum and maximum values of a parameter). Finally, the user uses `genExPrior()` to compute the ex-situ prior for a target site. The user has the options of using built-in plotting functions to visualize results.

The general workflow, which is followed in all of these examples, is visualized in Figure 4. First, a user has to represent the ex-situ data in the form of an `R dataframe` object. These data can be manually entered, like explained in Section 2.4.1, taken from the included data, like explained in Section 2.4.2, or being supplied though some additional database. Next, the user would enter as an `R vector` object a range of values at which to estimate the prior distribution. This range is typically the minimum and maximum values of the parameter of interest (for example, porosity takes values between 0 and 1). Finally, the user would input the ex-situ data and specified range into the function `genExPrior()`, which outputs a prior and the distributions of hyperparameters of the model. The user has the options of using built-in plotting functions to visualize results.

Example 1: Using exPrior with synthetic data

To familiarize the reader with this general workflow, let us start with a simple example using only a few synthetic data (on a log 10 scale) from three arbitrarily labeled sites S_1, S_2, and S_3. The source code for the corresponding vignette can be found at https://github.com/GeoStat-Bayesian/exPrior/blob/master/vignettes/using_genExPrior.Rmd. The goal is to derive the ex-situ prior for target site S_0 with the following code:

```
> exdata <- data.frame(val = c(c(-2,-3,-4), c(-2,-1), c(-6,-7,-2,-3)),
+                      site_id = c(rep('S1',3), rep('S2',2), rep('S3',4)))
> ex_prior <- genExPrior(exdata = exdata, theta = seq(from=-10, to=10, by=0.1))
```

By following the above workflow, we started with generating an `R dataframe` exdata for the ex-situ data. Then, we entered the range over which to estimate the variable θ as an `R vector` theta. The actual computation of the ex-situ prior is finally performed by the function `genExPrior()`. To investigate the output of this function, **exPrior** provides a number of plotting functions.

```
> plotHyperDist(ex_prior)
> plotExPrior(ex_prior)
```
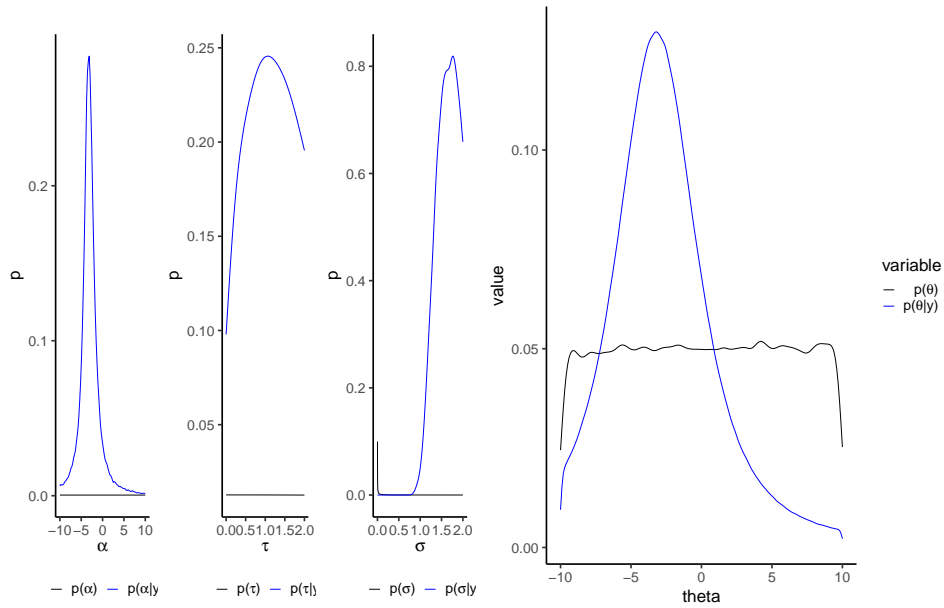
Figure 5: The left panel shows the distributions of the hyperparameters `alpha`, `tau`, and `sigma`. The right panel shows ex-situ prior computed using the data assimilation framework (blue curve) against the uninformative prior (black curve).

In this little example, the first command `plotHyperDist` plots the posterior distributions of the hyperparameters α, τ, and σ (see Figure 5 left panel). This captures the impact of the data on the Bayesian hierarchical model. The second command `plotExPrior` shows the ex-situ data from the three sites jointly with the predicted prior distribution for the new site S_0 (see Figure 5 right panel). As can be seen, the essentially flat, uninformative prior got updated into a much sharper, informative prior representing a much-reduced uncertainty.

Example 2: Using exPrior with real-world data

As introduced above, **exPrior** provides real-world, geostatistical data from the WWHYPDA. Let us exemplify its use and impact on the inference by first importing the data on porosity. As above, the associated vignette can be found at https://github.com/GeoStat-Bayesian/exPrior/blob/master/vignettes/real_world_data.Rmd.

```
> load(file="data/df_porosity.rda")
```

These real-world data are now loaded into the workspace and can be used to compute the ex-situ prior using the 'exPrior' function as describe in above example

```
> resExPrior = genExPrior(exdata = df_porosity, theta = seq(from=0, to=1, by=0.01))
```

Here, the range of the `theta` vector reflects the common-sense intuition that porosity values can only exist between 0 and 1. This change should also be reflected in the used model

$$y_{i,j} \sim \Phi(\mu_i, \sigma^2, 0, 1), \tag{4a}$$

$$\mu_i \sim \mathcal{N}(\alpha, \tau), \tag{4b}$$

$$(\sigma^2, \alpha, \tau) \sim p(\sigma^2)p(\alpha)p(\tau). \tag{4c}$$

Equation (4a) makes this change clear, such that the data are drawn from a truncated normal distribution. Since the boundaries are fixed, the hierarchical model itself still has the same number of parameters, and the other parts of the model remain the same.

After the completion of `exPrior`, we can visualize again the posteriors of the model as well as the prior of θ using `plotExPrior`.

```
> plotHyperDist(resExPrior)
> plotExPrior(resExPrior)
```
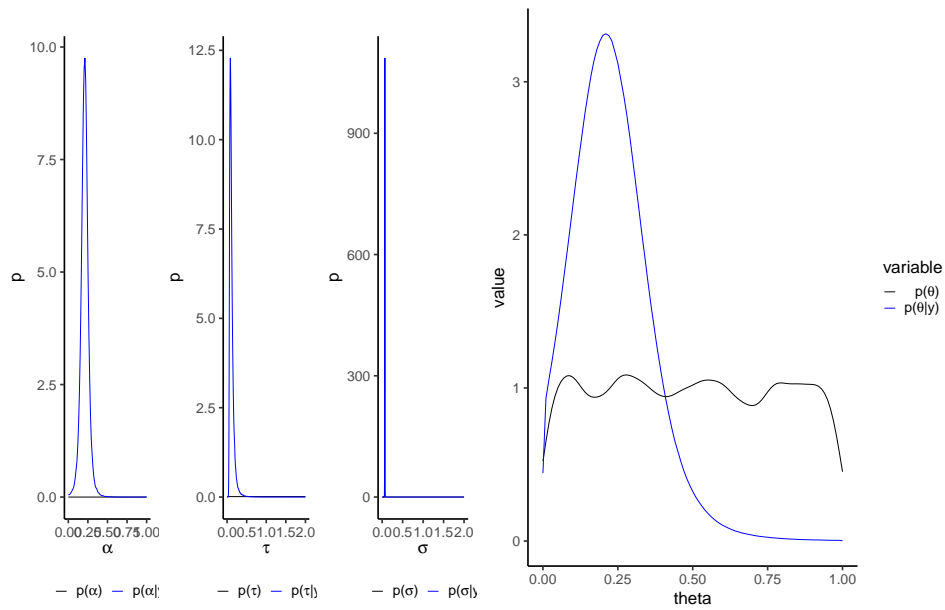
Figure 6: Informative (blue) and non-informative (black) priors computed with `genExPrior()` using real-world, ex-situ data of porosity in sand stone.

Compared to Figure 5, the results in Figure 6 show some relevant differences. In particular, the hyperpriors seen in the left panel of Figure 6 are much more peaked, resulting in near-certainty about their value. This is due to the large amount of evidence provided by the data. Since the parameter distributions on the higher levels in a hierarchical model represent the uncertainty about the parameters on the lower ones, it can be said that the results of this inference capture the uncertainty of porosity in sandstone with high certainty. This means that the prior distribution in the right panel of Figure 6 is very close to the statistical uncertainty for the hypothetical population of all porosity values in sandstone aquifers in general. As can be seen in this figure, this distribution is strongly peaked between 0.2 and 0.3. Using this prior therefore provides a practitioner with a sound foundation for the geostatistical inference of the in-situ porosity.

Example 3: Accounting for spatial autocorrelation in ex-situ data

In most cases, ex-situ data used in the analysis are spatially correlated since measurements are usually collected in a clustered way (Rubin, 2003; Pyrcz and Deutsch, 2002). The data assimilation model outlined in 2.2.1 can, in principle, account for patterns of spatial variability by using multivariate distributions as site-specific distributions. As above, the associated vignette can be found at https://github.com/GeoStat-Bayesian/exPrior/blob/master/vignettes/spatial_correlation.Rmd. To account for this spatial correlation, let us use a revised version of the hierarchical model from Equation (3)

$$y_{i,j} \sim \mathcal{N}(\mu_i, \Sigma), \tag{5a}$$

$$\mu_i \sim \mathcal{N}(\alpha, \tau), \tag{5b}$$

$$\Sigma = \sigma^2 \exp\left(-\frac{h}{\lambda}\right), \tag{5c}$$

$$(\sigma^2, \lambda, \alpha, \tau) \sim p(\sigma^2)p(\lambda)p(\alpha)p(\tau). \tag{5d}$$

The relevant adjustment can be seen in Equation (5a), where the data are no longer modeled to be drawn from a univariate normal distribution but a multivariate distribution instead. The main difference is the replacement of the variance σ^2 by the covariance Σ. In our example, this covariance is modeled as an isotropic exponentially decaying function, with a characteristic length scale λ. This function means that measurements being taken at large distances h are essentially independent, and no relevant difference to the simple univariate model from Equation (3) would exist. However, measurements taken at distances h similar or smaller to λ exhibit substantial correlation and must be
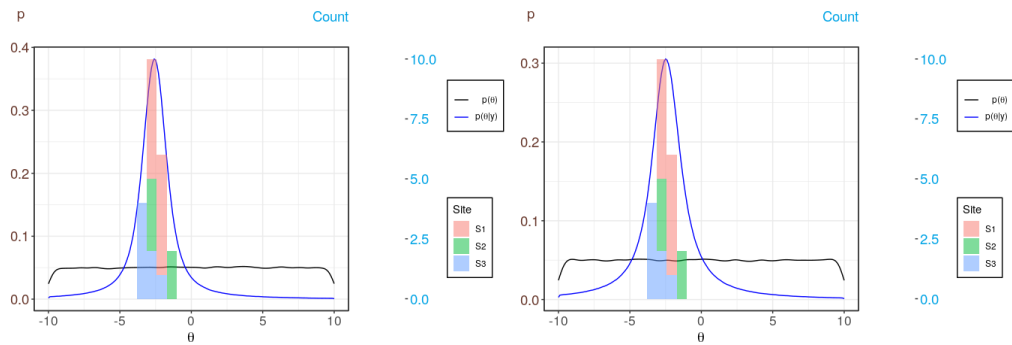
Figure 7: Informative (blue) and non-informative (black) priors computed with `genExPrior()`. The left panel shows the prior without accounting for spatial correlation, whereas the right panel shows the prior accounting for spatial correlation (right panel). The colored bars represent the ex-situ data from the three different sites.

assimilated accordingly. Failing to do so would result in an underestimation of the actual uncertainty, a phenomenon which is known in the literature as pseudoreplication (Hurlbert, 1984; Legendre, 1993). Due to the additional parameter, the model has now 4 hyperparameters, which need to be inferred.

To exemplify the workflow with this revised hierarchical model, let us use synthetic data coming again from three different sites only. To generate data with spatial correlation, we used the **gstat** package. These synthetic data were then transformed to have different mean values for each site.

```
> set.seed(1)
> xy <- data.frame("x" = sample(seq(0.00,1.00,0.01),22),
+                  "y" = sample(seq(0.00,1.00,0.01),22))
> model = vgm(psill=1, range=1, model='Exp')
> g.dummy <- gstat(formula=z~1, locations=~x+y, dummy=TRUE, beta=1, model=model, nmax=20)
> exdata_spatial <- predict(g.dummy, newdata=xy, nsim=1)
```

To adapt this data frame from **gstat** to the format needed for **exPrior**, we have to change one of the column names and add the site's id.

```
> colnames(exdata_spatial)[3] <- "val"
> exdata_spatial\$site_id = c(rep("S1", 10), rep("S2", 5), rep("S3", 7))
> exdata_spatial[ 1:10, 'val'] <- exdata_spatial[ 1:10, 'val'] - 3
> exdata_spatial[11:15, 'val'] <- exdata_spatial[11:15, 'val'] - 2.5
> exdata_spatial[16:22, 'val'] <- exdata_spatial[16:22, 'val'] - 3.5
```

With these data, we can now generate the ex-situ prior distribution. To tell **exPrior** to account for the spatial correlation in the data, we have to toggle the `spatialCoordinates` flag in the `genExPrior()` function to `TRUE`.

```
> resExPrior = genExPrior(exdata = exdata, theta = seq(from=-10, to=10, by=0.1),
+                         spatialCoordinates = TRUE)
> plotExPrior(resExPrior, plotExData = TRUE)
```

To compare the effects of accounting for spatial correlation, we provide plots of the ex-situ prior with `spatialCoordinates` being both set to `FALSE` and `TRUE` (see the left and right panel in Figure 7, respectively). As can be seen, both priors look overall similar in shape. The main difference is that the latter shows a somewhat increased uncertainty, i.e., a wider variance, which can be seen by the increased mode of the distribution. The fact that the more realistic model produces more uncertain results may seem counterintuitive at first. However, the aim of statistical inference is not to reduce the uncertainty as much as possible but to correctly capture the uncertainty in the used data and the model. As mentioned above, this problem of not accounting for possible correlations between measurements is called pseudoreplication and can have serious consequences by leading to overconfident statistical analyses.

Example 4: Assimilating Multiple Data Types

As mentioned above, **exPrior** is written in a flexible manner, such that it can assimilate data that come in the form of measurements, bounds, or moments (see schematic in Figure 2). To exemplify this
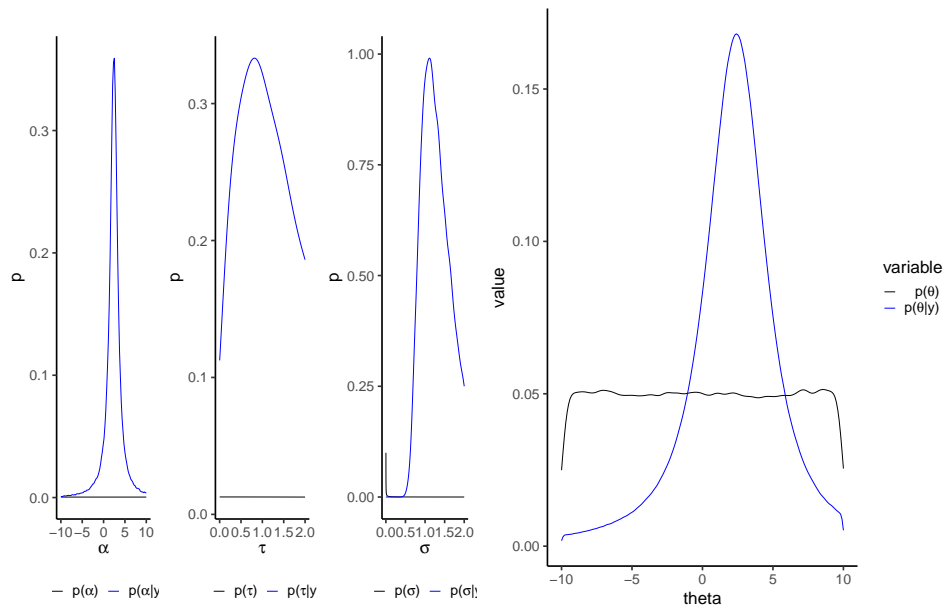
Figure 8: The left panel shows the distributions of the hyperparameters `alpha`, `tau`, and `sigma`. The right panel shows ex-situ data from three synthetic sites S_1, S_2, and S_3. The blue curve is the ex-situ prior computed using the data assimilation framework, while the black curve is the uninformative prior.

flexibility, let us use in this example synthetic data from three sites labeled S1, S2, and S3. From Site S1, we have data in the form of bounds, where the minimum value of a hydrogeological property of S1 is 2, and its maximum value is 4. Site S2 has data in the form of moments, where the first moment, or site mean, is 2, while the second moment, or site variance, is 0.1. Finally, site S3 has three measurements. Again, the associated vignette can be found at https://github.com/GeoStat-Bayesian/exPrior/blob/master/vignettes/multi_type_data.Rmd. The code below shows how to format the data in R such that it can be read into genExPrior().

```
> exdata_S1 <- data.frame(val=c(2,4), site_id=rep('S1',2),
+                         type=c('bound.min','bound.max'))
> exdata_S2 <- data.frame(val=c(2,0.1), site_id=rep('S2',2),
+                         type=c('moment.1','moment.2'))
> exdata_S3 <- data.frame(val=c(2,3,4), site_id=rep('S3',3),
+                         type=c('meas','meas','meas'))
> exdata <- rbind(exdata_S1, exdata_S2, exdata_S3)
```

As in previous examples, the data frame `exdata_multitype` as well as the vector `theta` can be input directly into `genExPrior()` as such

```
> resExPrior <- genExPrior(exdata = exdata, theta = seq(from=-10, to=10, by=0.1))
```

Finally, we can visualize the results `resExPrior` again using the `plotHyperDist` and `plotExPrior` functions.

```
> plotHyperDist(resExPrior)
> plotExPrior(resExPrior)
```

The resulting hyperparameters and ex-situ prior distributions look very similar to the simple example from Section 2.4.1 (compare Figure 5 to Figure 8). This comparison shows that data in the form of bounds and moments can have a similar impact on the inference and how they can be assimilated by `exPrior`.

Summary

In this paper, we have introduced the R package **exPrior**, which contains methods for assimilating ex-situ data to generate prior probabilities for geostatistical parameters. We explain the formulation

of a prior distribution as a Bayesian Hierarchical Model (Section 2.2.1) and its implementation using **NIMBLE**, an R package created for efficient hierarchical modeling. We illustrate the model through a number of examples where **exPrior** can be used, including univariate and multivariate models (Section 2.4.3), as well as the assimilation of multiple data types (2.4.4). The package also contains data from the WWHYPDA, an open-source, hydrogeological database that provides valuable information for hydrogeological modeling. The goal of this package is to provide methods to facilitate geostatistical modeling, as well as to encourage the open-source and open-data movements between scientists.

Acknowledgements

Bibliography

K. S. Bakar and S. K. Sahu. spTimer: Spatio-Temporal Bayesian Modeling Using R. *Journal of Statistical Software*, 63(15):32, 2015. ISSN 1548-7660. doi: 10.18637/jss.v063.i15. [p101]

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, 2nd edition, Sep 2014. [p101]

D. Beaudette, P. Roudier, and A. O'Geen. Algorithms for quantitative pedology: A toolkit for soil scientists. *Computers & Geosciences*, 52:258 – 268, 2013. ISSN 0098-3004. doi: https://doi.org/10.1016/j.cageo.2012.10.020. [p101]

J. Berger. The Case for Objective Bayesian Analysis. *Bayesian Analysis*, 1(3):385–402, 2006. ISSN 1931-6690. doi: 10.1214/06-BA115. [p102]

A. Comunian and P. Renard. Introducing wwhypda: a world-wide collaborative hydrogeological parameters database. *Hydrogeology Journal*, 17(2):481–489, 2009. ISSN 1435-0157. doi: 10.1007/s10040-008-0387-x. [p101, 102, 106]

K. Cucchi, F. Heße, N. Kawa, C. Wang, and Y. Rubin. Ex-situ priors: A Bayesian hierarchical framework for defining informative prior distributions in hydrogeology. *Advances in Water Resources*, 126:65 – 78, 2019. ISSN 0309-1708. doi: 10.1016/j.advwatres.2019.02.003. [p101]

P. de Valpine, D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik. Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. *Journal of Computational and Graphical Statistics*, 26(2):403–413, 2017. doi: 10.1080/10618600.2016.1172487. [p101]

P. J. Diggle and P. J. Ribeiro. *Model-based Geostatistics*. Springer, 1rst edition, 2007. [p101]

A. O. Finley, S. Banerjee, and A. E. Gelfand. spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models. *Journal of Statistical Software*, 63(13):1–28, 2015. ISSN 1548-7660. doi: 10.18637/jss.v063.i13. [p101]

A. E. Gelfand. Hierarchical modeling for spatial data problems. *Spatial Statistics*, 1(Supplement C):30 – 39, 2012. ISSN 2211-6753. doi: 10.1016/j.spasta.2012.02.005. [p102]

A. E. Gelfand and E. M. Schliep. Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics*, 18(A, SI):86–104, Nov 2016. ISSN 2211-6753. doi: 10.1016/j.spasta.2016.03.006. [p102]

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science (Book 106). Chapman and Hall/CRC, 3 edition, 2013. [p102, 103]

F. Heße, A. Comunian, and S. Attinger. What we talk about when we talk about uncertainty. Toward a unified, data-driven framework for uncertainty characterization in hydrogeology. *Frontiers in Earth Science*, 2019a. doi: 10.3389/feart.2019.00118. [p101]

F. Heße, K. Cucchi, and N. Kawa. Geostat-bayesian/exprior: First release, Nov. 2019b. URL https://doi.org/10.5281/zenodo.3544517. [p101]

F. Heße, K. Cucchi, and N. Kawa. Geostat-bayesian/geostatdb: First release, Oct. 2019c. URL https://doi.org/10.5281/zenodo.3473022. [p102, 106]

R. J. Hoeksema and P. K. Kitanidis. Analysis of the spatial structure of properties of selected aquifers. *Water Resources Research*, 21(4):563–572, 1985. ISSN 1944-7973. doi: 10.1029/WR021i004p00563. [p103]

S. H. Hurlbert. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54(2):187–211, 1984. ISSN 00129615. URL http://www.jstor.org/stable/1942661. [p111]

N. Kawa, K. Cucchi, and F. Heße. Geostat-bayesian/sitesimilarity: First release, Sept. 2020. URL https://doi.org/10.5281/zenodo.4059706. [p102]

P. Kitanidis. *Introduction to Geostatistics: Applications in Hydrogeology*. Cambridge University Press, 2008. [p101, 102]

J. K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press, 2010. [p102]

P. Legendre. Spatial Autocorrelation: Trouble or New Paradigm? *Ecology*, 74(6):1659–1673, Sep 1993. ISSN 0012-9658. doi: {10.2307/1939924}. [p111]

F. Lindgren and H. Rue. Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, 63(19): 1–25, Jan 2015. ISSN 1548-7660. doi: 10.18637/jss.v063.i19. [p101]

D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The bugs project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067, 2009. ISSN 1097-0258. doi: 10.1002/sim.3680. [p101]

MINES ParisTech / ARMINES. RGeostats: The Geostatistical R Package. Free download from: http://cg.ensmp.fr/rgeostats, 2019. [p101]

A. Papritz, C. Schwierz, and M. Papritz. georob. https://CRAN.R-project.org/package=georob, 2014. [p101]

E. J. Pebesma. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30(7):683 – 691, 2004. ISSN 0098-3004. doi: 10.1016/j.cageo.2004.03.012. [p101, 102]

M. J. Pyrcz and C. V. Deutsch. *Geostatistical Reservoir Modeling*. Oxford University Press, 2002. [p101, 102, 110]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/. [p101]

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. [p102]

P. J. Ribeiro and P. J. Diggle. geoR: a package for geostatistical analysis. *R-News*, 1:15–18, 2001. [p101]

M. D. Risser and D. Turek. Bayesian inference for high-dimensional nonstationary gaussian processes, 2020. [p101]

Y. Rubin. *Applied Stochastic Hydrogeology*. Oxford University Press, USA, 2003. [p101, 102, 110]

Y. Rubin, X. Chen, H. Murakami, and M. Hahn. A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields. *Water Resources Research*, 46, Oct 6 2010. ISSN 0043-1397. doi: {10.1029/2009WR008799}. [p101]

Y. Rubin, C.-F. Chang, J. Chen, K. Cucchi, B. Harken, F. Heße, and H. Savoy. Stochastic hydrogeology's biggest hurdles analyzed and its big blind spot. *Hydrology and Earth System Sciences Discussions*, 2018:1–36, 2018. doi: 10.5194/hess-2018-290. [p101]

H. Savoy, F. Heße, and Y. Rubin. anchoredDistr: a Package for the Bayesian Inversion of Geostatistical Parameters with Multi-type and Multi-scale Data. *The R Journal*, 9(2):6–17, 2017. doi: 10.32614/RJ-2017-034. [p101]

L. J. Slater, G. Thirel, S. Harrigan, O. Delaigue, A. Hurley, A. Khouakhi, I. Prosdocimi, C. Vitolo, and K. Smith. Using R in hydrology: a review of recent developments and future directions. *Hydrology and Earth System Sciences*, 23(7):2939–2963, 2019. doi: 10.5194/hess-23-2939-2019. URL https://www.hydrol-earth-syst-sci.net/23/2939/2019/. [p101]

J. Vrugt, C. ter Braak, H. Gupta, and B. Robinson. Equifinality of formal (dream) and informal (glue) bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, 23(7):1011–1026, 2009. ISSN 1436-3240. doi: 10.1007/s00477-008-0274-y. [p101]

Falk Heße
Institute of Earth and Environmental Sciences
University Potsdam
Potsdam, Germany
ORCiD: 0000-0002-2547-8102
falk.hesse@ufz.de

Karina Cucchi
Department of Civil and Environmental Engineering
University of California, Berkeley
Berkeley, CA, USA
karina.cucchi@berkeley.edu

Nura Kawa
Department of Statistics
University of California, Berkeley
Berkeley, CA, USA
currently at:
Leuven Statistics Research Centre,
KU Leuven,
Leuven, Belgium
nkawa@berkeley.edu

Yoram Rubin
Department of Civil and Environmental Engineering
University of California, Berkeley
Berkeley, CA, USA
rubin@ce.berkeley.edu