Groundwater

Methods Note/

Defining Hydrogeological Site Similarity with Hierarchical Agglomerative Clustering

by Nura Kawa[1,2], Karina Cucchi[3], Yoram Rubin[3], Sabine Attinger[4,5], and Falk Heße[6,7]

Abstract

Hydrogeological information about an aquifer is difficult and costly to obtain, yet essential for the efficient management of groundwater resources. Transferring information from sampled sites to a specific site of interest can provide information when site-specific data is lacking. Central to this approach is the notion of site similarity, which is necessary for determining relevant sites to include in the data transfer process. In this paper, we present a data-driven method for defining site similarity. We apply this method to selecting groups of similar sites from which to derive prior distributions for the Bayesian estimation of hydraulic conductivity measurements at sites of interest. We conclude that there is now a unique opportunity to combine hydrogeological expertise with data-driven methods to improve the predictive ability of stochastic hydrogeological models.

Introduction

A good understanding and accurate description of subsurface conditions of a hydrogeological site is important for a variety of applications. Examples include water management for freshwater supply, oil production, CO_2 sequestration, and modeling the transport of contaminants in the subsurface. Unfortunately, knowledge of such conditions can be highly uncertain due to the heterogeneity of subsurface properties such as hydraulic conductivity and porosity. Further complicating matters, standard subsurface exploration techniques are challenging and costly. As a result, limited data are typically available to characterize a given site, making prediction uncertainty very high.

Under such circumstance of data scarcity, practitioners should incorporate all available data sources on a given site to reduce the uncertainty as much as possible (Rubin et al. 2018). Bayesian methods have been proven to provide a framework wherein heterogeneous data sources can be joined to represent the available knowledge of a given situation (Heße et al. 2019a). This is achieved by distinguishing between two different sources of data: case-specific data and background data, which are then combined into a full representation using Bayes' theorem (Kruschke 2010; Gelman et al. 2013). Case-specific data, which in the case of hydrogeology would be in situ data, are represented through the likelihood, whereas available background knowledge is represented through the prior distribution (Ulrych et al. 2001; Gelman 2006). The use of Bayesian methods has drawn criticism, specifically regarding the choice of prior distribution (Easwaran 2011a, 2011b). The choice

[1]Department of Statistics, University of California Berkeley, Berkeley, CA
[2]Department of Mathematics, Leuven Statistics Research Centre, KU Leuven, Leuven, Belgium
[3]Department of Civil and Environmental Engineering, University of California Berkeley, Berkeley, CA
[4]Department of Geosciences, University of Potsdam, Potsdam, Germany
[5]Department of Computational Hydrosystems, Helmholtz Center for Environmental Research—UFZ, Leipzig, Germany
[6]Corresponding author: Department of Geosciences, University of Potsdam, Potsdam, Germany; falk.hesse@ufz.de
[7]Department of Computational Hydrosystems, Helmholtz Center for Environmental Research—UFZ, Leipzig, Germany

Article impact statement: This article introduces hierarchical clustering as a method for defining a notion of site similarity; the aim of this method is to improve the derivation of prior distributions in Bayesian methods in hydrogeology.

Received August 2021, accepted June 2022.

doi: 10.1111/gwat.13261

of prior distribution impacts the final result of Bayesian inference; a poorly chosen prior can reduce the accuracy of the resulting posterior. There exists no single consensus on how to define and select a prior distribution, or how to properly represent a lack of knowledge via diffuse priors.

We propose to use *informed prior distributions*, that is, priors that are derived from relevant background knowledge, which reflect a state of partial certainty. In addition to avoiding the negative impact of a poorly chosen prior on the final result, the use of informed priors has other benefits. In a data-scarce context, where the parameters of the likelihood are estimated using very few measurements, an informed prior can yield more accurate predictions of site characteristics than can the use of flat priors. Depending on its form, an informed prior can also help researchers choose the most appropriate exploration technique and guide its deployment. Overall, this can reduce exploration costs and help determine appropriate experimental design.

Cucchi et al. (2019) introduced the *exPrior* algorithm to derive informed priors. The algorithm fits a Bayesian hierarchical model with measurements obtained from relevant sites, yielding a posterior predictive distribution that can be used as a prior distribution for an unexplored site. This algorithm requires a user to provide hydrogological measurements in order to derive an informed prior. In data-scarce scenarios, a user can obtain data from an open-source database: the World-Wide HYdrogeological Parameters DAtabase (WWHYPDA) (Comunian and Renard 2009). Both the *exPrior* algorithm and functions to load and query the WWHYPDA can be accessed in the `R` programming language: The algorithm is implemented in package `exPrior` (Heße et al. 2021), and the database can be accessed using `geostatDB` (Heße et al. 2019c), an `R` interface for accessing hydrogeological information from the WWHYPDA.

Cucchi et al. (2019) recommends to derive informed priors with *exPrior* algorithm using data from sites that are similar to the site of interest, which we call the *target site*. Similarity is defined on the basis of physical characteristics, which refer to the state of the physical system under consideration, as opposed to epistemic characteristics which refer to how the observer interacts with the physical system to generate information. (Cucchi et al. 2019). Ultimately, the selection of similar sites is left to domain experts.

In theory, using as much data as possible to derive a prior, or assimilating all measurements from roughly similar sites, could result in an informed prior with large uncertainty, perhaps not much better than a flat prior. Using instead measurements obtained from sites that are very similar to the target site would yield a more peaked, relevant informed prior. However, there may exist sites where little to no data from similar sites is available to inform a prior distribution. Clearly, a good compromise is needed when pooling data from a given database (Halpern 2003; Jaeger 2006). In statistics, this is known as the *reference class problem*; in other words, that there is no unique reference class from which a given object could be considered to belong (Hajek 2007; Hajek and Hitchcock 2016; Wallmann 2017). Within the context of hydrogeology, this means it is not clear how to decide from which sites to transfer data. Is a target site most like sites with similar rock types? Or is it more similar to sites from the same type of environment or region?

Such a procedure of using only data from a limited number of similar sites is quite common in the related field of hydrology. In this domain, criteria used to select sites include physical proximity (Merz and Blöschl 2005) and climatic and physiographic properties (Blöschl and Sivapalan 1995; McIntyre et al. 2005). Additionally, one can use empirical transfer functions (Zacharias and Wessolek 2007; Kumar et al. 2013) or rely on expert knowledge Li et al. (2017).

We propose an alternative approach: to use machine learning to determine which sites to use to derive informed priors. Algorithmic selection could ensure reproducibility, as well as encourage the evolution of the notion of site similarity with the increased availability of data. In particular, we suggest Hierarchical Agglomerative Clustering, which partitions a dataset into similarity-based groups. This algorithm has been successfully applied in other fields to determine groups of similar objects in a dataset.

The paper is organized as follows: first, we introduce our method for determining a set of similar sites for a given target site, using the *WWHYPDA* as a reference database. We then outline a use case for our method: to use it as a preprocessing step for deriving informative prior distributions. We describe an experiment to determine whether our method improves the quality of informative priors derived with the *exPrior* algorithm, again using data from the *WWHYPDA*. In the discussion, we include explanations of scenarios where our method for deriving informed prior distributions with similar sites did not produce better results than using a prior derived from all available information. Finally, we outline directions for further research and improvement of our method.

Research Method

In the following, we present a data-driven method for defining site similarity. Using the *WWHYPDA* as a reference database, we represent hydrogeological sites as feature vectors, then use clustering to group them by similarity. We then apply the method as a data selection step for deriving informative prior distributions of hydraulic conductivity for 52 sites in the *WWHYPDA*.

Data: The *WWHYPDA*

To demonstrate our method, we use the WorldWide HYdrogeological Parameters DAtabase (*WWHYPDA*) introduced by Comunian and Renard (2009), the largest open source database of hydrogeological measurements. This database is designed to store values of the most important properties of earth materials, supplementing hydrogeological studies with additional data. The tabular database has an entity-relationship schema, where a basic entity is a sample of measurements taken at

a site. Additional information about the site exists in separate tables, with which a user can link to a measurement. In particular, the database relates each sample of measurements to a hydrogeological site to an earth material (rock type) and to a hydrogeological environment (environment type). Rocks and environment types are presented as tree structures, where instances are organized into parent and child relationships. The base elements (parents) correspond to most common families, the sub elements are refinements in the classification. We leave a more detailed description of the content and structure of the database to Comunian and Renard (2009).

The *WWHYPDA* is, within knowledge of the authors, the largest database of hydrogeological parameters. Currently, it contains a total of 20,523 measurements of 6 hydrogeological parameters spanning 128 sites. Additionally, the *WWHYPDA* is an open source and open-access database, which a user can query with SQL without barriers. In this paper we focus on hydraulic conductivity; the most common measurement type in the *WWHYPDA*.

For the purpose of demonstrating our method, we extracted from all 12,505 measurements of hydraulic conductivity, and for each datum we obtained the following features:

- *site name*: the name of the site where the measurement was taken,
- *rock type*: the corresponding rock type of the measurement,
- *parent rock type*: the family to which the rock type belongs,
- *environment type*: the corresponding hydrogeological environment of the measurement (e.g. "Sedimentary environment," "Volcanic environment"),
- *parent environment type*: the family to which the environment belongs,
- *fracturation degree*: the fracturation degree at the location of the measurement.

We use these six categorical features, or attributes, to characterize hydrogological sites in the database. The data in our method is therefore organized by site; a basic entity is a site name, and for each site we have a set of measurements of hydraulic conductivity, as well as six categorical features.

Hierarchical Agglomerative Clustering Using Categorical Feature Data

Hierarchical agglomerative clustering (HAC) is among the most established approaches to clustering data (Bandyopadhyay and Saha 2013). The algorithm partitions data into groups based on the similarity between observations (in our study between sites), computed mathematically as a distance. This approach is best suited for clustering observations with an already existing hierarchical structure (Hastie et al. 2009), such as the parent child categorization of earth materials in the *WWHYPDA* (see previous Section Data: The WWHYPDA) and other geological and hydrogeological classification schema. To familiarize the reader with HAC, let us briefly describe the algorithm in three steps.

Step One: HAC Input Data and Proximity Matrix

We begin with a set S of observations, $S = \{v_1, v_2, \ldots, v_N\}$ that we wish to partition into clusters. Here, each v_i is a vector describing the attributes of one hydrogeological site. To create a dataset suitable for clustering, we first arrange S into a row wise observation attribute matrix, where each v_i becomes a matrix row, and attributes of the v_i are the matrix columns. Next, we compute a proximity matrix comprised of the pairwise distances between rows of the observation attribute matrix. This becomes the input data for HAC.

The distance metric used should match the type(s) of data present in the v_i. A standard distance used is the Euclidean distance, which measures distance between metric data. For categorical or mixed-type data, one can use Gower's distance. In our experiment, we convert the categorical features obtained from the *WWHYPDA* into a set of binary v_i. Therefore, we use the Jaccard distance d_J as the distance metric between two observations. The Jaccard distance between observations v_i and v_j is defined as follows (Lung and Zhou 2010):

$$d_J(v_i, v_j) = 1 - J(v_i, v_j),$$
$$J(v_i, v_j) = \frac{M_{1,1}}{M_{1,1} + M_{0,1} + M_{1,0}}, \quad (1)$$

where $M_{1,1}$, $M_{1,0}$, $M_{0,1}$ are counts of 1-1, 1-0, and 0-1 matches of attribute pair between v_i and v_j. J is called the Jaccard similarity coefficient and measures the proportion of overlapping positive attributes among non-null attributes. Similar observations have a Jaccard similarity coefficient close to 1 and a Jaccard distance close to 0. Other usable distance metrics for binary observations include the simple matching coefficient and cosine distance. However, the Jaccard distance is recommended when working with binary variables with unequal frequencies of 0 and 1 (Bandyopadhyay and Saha 2013). This is because the Jaccard coefficient counts only present attributes, which, in our data, are more indicative of site similarity than are absent attributes.

We compute a set of N distances for each of the N vectors v_i and arrange them in a proximity matrix, D. D is a square matrix with dimension $N \times N$, where each entry represents a distance between observations. Specifically, each entry of D is the distance between observations i and j, where $D[i, j] = d_J(v_i, v_j)$, and $D[i, j] = 0$ for $i = j$. Note that $d_J(.)$ can be replaced by any appropriate distance or dissimilarity metric.

Step Two: HAC Algorithm

HAC is a *bottom-up* clustering algorithm, treating each observation as its own singleton cluster (singleton refers to a group with only one element). At each iteration, HAC merges the two clusters with the smallest group distance, computed from proximity matrix D. The algorithm stops when all groups have been merged into one large

cluster. The definition of group distance, called a linkage method, depends on the variation of the algorithm. Single linkage, or the nearest neighbor method, defines group distance as the smallest distance between observations belonging to each cluster. Complete linkage, or farthest neighbor method, defines cluster distance as that of the largest distance between observations in each cluster. Average linkage is a compromise between the former two. Single linkage often creates HAC results with "chaining," or a series of many singleton clusters merged at higher levels. In contrast, complete linkage creates more "round" clusters that are well defined (Hastie et al. 2009). Thus, in this paper, we use complete linkage. An HAC clustering is visualized as a dendrogram, a tree of observations connected by horizontal lines, which represent merges that the algorithm performs. The height of a dendrogram measures the proximity between clusters, which, for Jaccard distance, ranges from 0 to 1. The `stats` package in the R Programming language has an implementation of HAC with complete linkage (R Core Team 2017).

Step Three: Cutting the HAC Tree to Decide k

Finally, we cut the HAC dendrogram horizontally at a height that produces the k number of clusters such that the resulting partition best satisfies our criterion for clustering: that clusters have a small average within-cluster distance. To determine k we use a selection method described in Manning et al. (2008), which is to cut a dendrogram at multiple k and for each cut compute $W(k)$, the average within-cluster distance between observations. For a cut into k groups, we define the average within-cluster proximity as:

$$W(k) = \frac{1}{k}\sum_{i=1}^{k} d_{J_{C_i}}, \tag{2}$$

where $d_{J_{C_i}}$ is the average of the pairwise Jaccard distances of observations in a cluster C_i. Clusters with one observation are assigned a value of 0. Naturally, a good partition has a small $W(k)$. However, on average $W(k)$ decreases with k, especially because the number of singleton clusters increases with k. Therefore, we could select $k*$, the "optimal" k, to be that for which the gap between two successive $W(k)$ is largest. This method is commonly referred to as the *elbow rule*: one visually plots a set of $W(k)$ against their successive k and selects the $k*$ at the elbow of the graph.

However, for the application of predicting site qualities, we want to avoid singleton clusters, as they provide no similar sites to a target site. Therefore, we add to $W(k)$ a penalty for singletons:

$$W_p(k) = \frac{1}{k}\sum_{i=1}^{k}\left(d_{J_{C_i}} + \mathbb{1}\left(|C_i| = 1\right)\right), \tag{3}$$

where $\mathbb{1}\left(|C_i| = 1\right)$ is an indicator that takes value 1 if the cardinality of cluster C_i is equal to 1 (i.e., C_i is a singleton cluster) and 0 otherwise. We select $k*$ to be the smallest k for which $W_p(k)$ is minimized.

Validation of Clustering: The Silhouette

Well separated clusters have, on average, small within cluster dissimilarity and a large between cluster dissimilarity. The *silhouette* (Rousseeuw 1987) of an object measures the ratio between its within cluster similarity and its similarity to its nearest neighboring cluster. Using the Jaccard distance to measure dissimilarity between objects, the silhouette s of an object i is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{4}$$

where $a(i)$ is the average within cluster dissimilarity and $b(i)$ is the smallest average dissimilarity between i and members of a different cluster. Silhouettes range from -1 to 1, where well partitioned objects have silhouettes close to 1, misplaced objects have a negative value, and objects that lie between clusters have a value close to 0. To evaluate the quality of a cluster, we measure its average silhouette width, which is the average of the silhouettes of its individual members.

Application: Predicting Hydrogeological Properties at Unsampled Sites

Clustering results in a grouping of sites by their observable features (rock type, environment type, and fracturation degree). We evaluate the relevance of using the obtained groups as a reference for transferring earth material properties across hydrogeological sites. The underlying assumption is that the prediction of earth material properties at a new site is improved when limiting assimilated data to sites with similar observable features only. In this section, we describe the method used to predict hydraulic conductivity at one site based on measurements at similar sites, and the steps used to evaluate the extent to which clustering improves prediction.

Predicting Hydraulic Conductivity from Measurements at Similar Sites: Cucchi et al. (2019) introduced a Bayesian data assimilation framework for predicting earth material properties at an unsampled site, or *target site*, based on data from similar sites, or *reference sites*. Data assimilation is performed in a two-step algorithm that produces an informative pdf for a hydrological property of interest at the target site, called ex situ *prior pdf*. This ex situ prior summarizes information available about the property of interest at the target site and can be used as a starting point for further investigation. In this paper, the ex situ prior of interest is $p(Y|\mathcal{D})$, where Y is the random variable for log-transformed hydraulic conductivity and $\mathcal{D}$ are the ex situ data used in the assimilation. Here, $\mathcal{D} = y_{i,j}$, where $y_{i,j}$ is the log-transform of hydraulic conductivity measurement j at similar site i. The data assimilation framework fits the hierarchical model of the form

$$\begin{aligned} y_{i,j} &\sim N\left(\mu_i, \sigma^2\right) \\ \mu_i &\sim N\left(\alpha, \tau^2\right) \end{aligned} \tag{5}$$

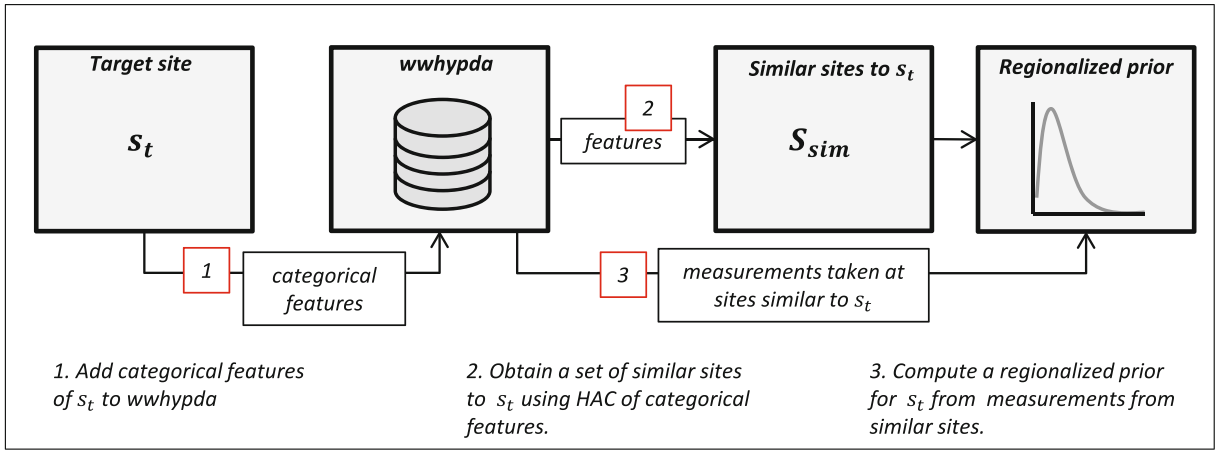

Figure 1. The procedure for computing a ex situ prior of the distribution of a parameter Y for a target site s_t.

where μ_i is the site-specific mean at site i, σ^2 represents within-site variability and τ^2 represents inter site variability.

Following the Bayesian framework, hyperparameters α, τ, and σ are given noninformative priors and their distributions are updated using Markov chain Monte Carlo sampling (Cucchi et al. 2019). The ex situ prior is the posterior predictive distribution for Y derived by marginalization over updated distributions of hyperparameters (Gelman et al. 2014). Methods for fitting the model to provided data and for deriving $p(Y|\mathcal{D})$ are provided within the `exPrior` R package, available on CRAN as well as GitHub (Heße et al. 2019b). In this paper, we introduce the use of clustering to determine for a target site a set of similar sites whose measurements are used as ex situ data $\mathcal{D}$. A general procedure for computing an ex situ prior of parameter Y at the target site is visualized in Figure 1.

For details, we refer to Cucchi et al. (2019) as well as the documentation on the project's GitHub page (Heße et al. 2019c).

Validation of Prediction Accuracy: Clustering provides a natural framework for the selection of reference sites to an unsampled target site s_t. Sites belonging to the same cluster as s_t become reference sites whose measurements of Y are used to compute an ex situ prior for s_t. Such priors can be more predictive of the actual statistical distribution of Y at s_t than are priors, computed from measurements sampled at all available sites, regardless of proximity to s_t. In the following, we will call the former *regionalized priors*. This nomenclature is borrowed from a similar procedure in hydrological modeling where it is used to describe the transfer of parameters from calibrated, donor catchments to an ungauged, target catchment. To test the accuracy, we compute regionalized and nonregionalized priors for sites where we have samples. Here, *nonregionalized priors* are prior distributions computed from all available measurements, regardless of relevance to s_t. We determine the prediction accuracy of a prior by comparing its shape and location to a parametric estimation of the distribution of measurements in site s_t, which we assume to be Normally distributed with mean and variance computed from measurements Y taken at s_t. We call this the *target distribution*. Alternatively, one can use a kernel density estimate of measurements of Y taken at s_t. A good prior distribution is as close as possible to a distribution estimated using actual measurements of s_t. To measure prediction accuracy, we look at (1) the difference between the distributions' medians (which, for Normal distributions, is the same as the mean and mode) and (2) the Kullback Leibler Divergence (KLD), used by Tang et al. (2016) to measure information from a prior, between the prior and the target distribution.

To measure the difference in location of a prior and a parametric estimate, we compute the absolute difference between their medians. Similar distributions are located in close proximity. To mathematically determine the difference between two PDFs, we use the Kullback-Leibler divergence (KLD). For a parameter Y that is a continuous random variable, the KLD is defined as:

$$d_{KL}(p(Y), q(Y)) = \int_{-\infty}^{\infty} p(Y) \ln\left(\frac{p(Y)}{q(Y)}\right) \mathrm{d}Y, \qquad (6)$$

where $p(Y)$ and $q(Y)$ are density functions of Y. In this context, $p(Y)$ denotes the target density, and $q(Y)$ is an estimated density of Y. The KLD measures how much information is lost if we use the estimated density instead of the actual density of Y. Densities with closer proximity have a smaller KLD, with 0 being exact similarity.

To assess the value gained from using a regionalized prior instead of a nonregionalized prior, we compare (1) the KLD from the target distribution and (2) the difference in median location from the target distribution. If a regionalized prior outperforms a nonregionalized prior, then it will have a smaller divergence from the target distribution, and its median will be located closer to that of the target distribution. Therefore, the difference between nonregionalized and regionalized priors is positive. If not, then the opposite will be true, that is, the regionalized prior will be closer in shape and location to the target distribution; the difference in value is negative.

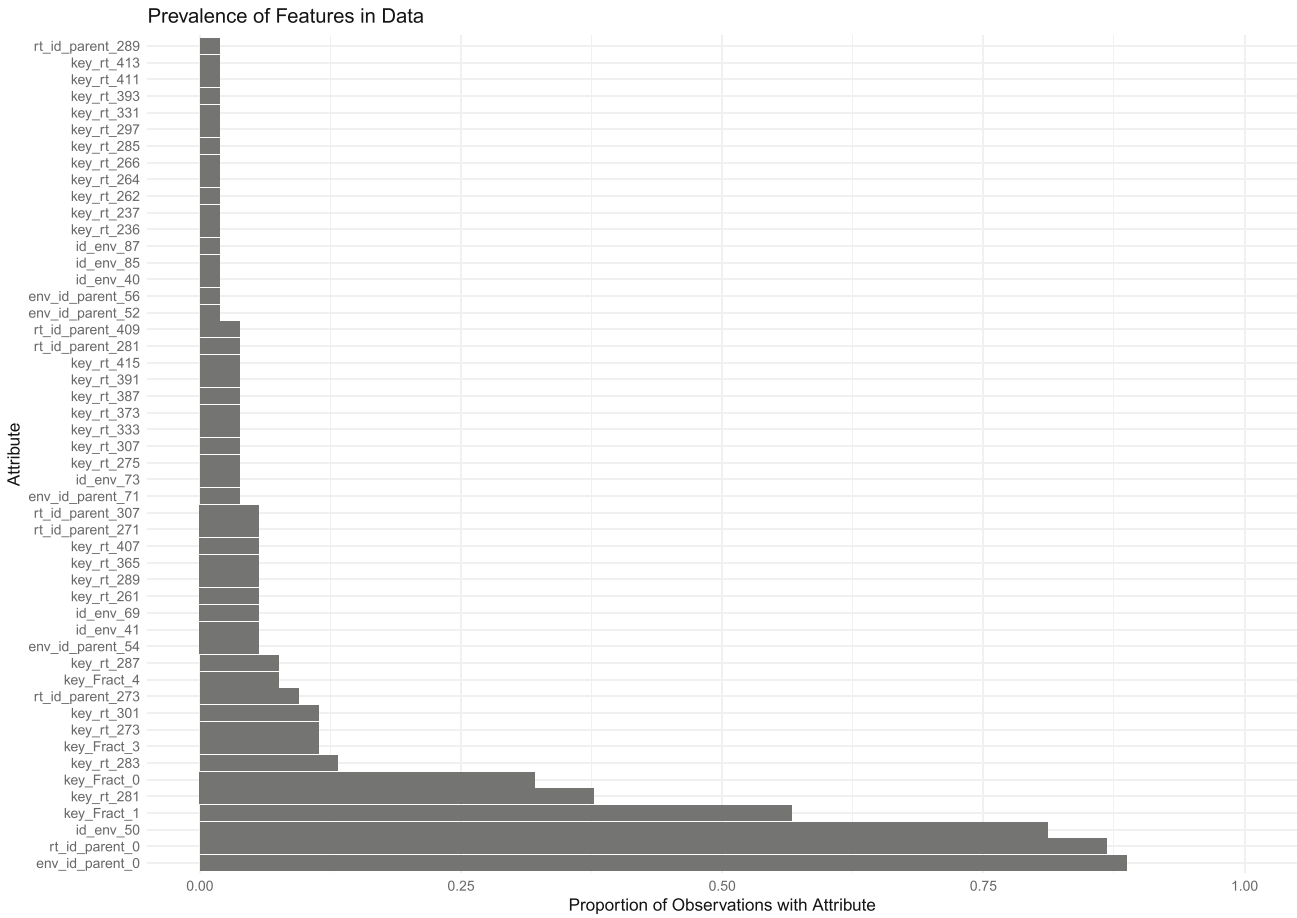

Figure 2. Frequency of observable attributes of hydrogeological sites used in clustering. Most common are *rock type parent 0* (88%), *environment parent 0—generic hydrogeological environment* (88%), *environment type 50—sedimentary environment* (82%), *fracturation degree 1—low fracturation* (57%) and *rock type 281—Sandstone* (41%). The median attribute frequency is 4%.

Results

We extracted a subset of the hydrogeological dataset, where we only considered sites where at least two measurements of hydraulic conductivity were taken. This left us with a set of 53 sites with 12,495 total measurements of hydraulic conductivity, where the number of measurements per site ranged from 2 to 3072. For each site we created a categorical vector of its features: *rock type*, *rock type parent*, *environment type*, *environment type parent*, and *fracturation degree*. Sites can be matched with multiple instances of rock type and environment type, for example, one site can be characterized with soil types "sand" and "gravel" and environment type "fluvial environment." Therefore, we transformed the categorical vectors into a set of 53 binary vectors, where each binary vector indicated the presence (denoted as 1) or absence (denoted as 0) of one single instance of a hydrogeological feature. We then arranged these binary vectors into the row wise observation attribute matrix, where rows represented sites and columns represented site attributes.

Let us start by giving an overview of the dataset used for this study, namely the attributes or features contained in the dataset. The frequency of attributes is depicted in Figure 2. As can be seen, the five most frequent attributes are "rock type parent 0", "environmental parent 0 (generic environment)", "environment type 50 (sedimentary environment)", "fracturation degree 1 (low)" and "rock type 281 (Sandstone)"; they are present in over 40% of hydrogeological sites in the database. The median frequency of an attribute is 4%, meaning that half of the attributes appear in less than 4% of sites. Additionally, 44 of 50 attributes have a frequency of less than 20%. While this distribution of attribute frequency is not problematic for the application of the HAC algorithm, it indicates that modifying the distance metric used for clustering by weighting attributes by their frequency could better distinguish sites.

Clustering: Having presented the data themselves, let us now turn our attention to the process of clustering the sites. As described above, we used the observation attribute matrix with HAC using complete linkage.

To visualizes the process for selecting the number of clusters, we show a dendrogram in Figure 3. A dendrogram is a tree diagram suitable for showing hierarchical associations. Each observation begins at the bottom of the dendrogram with height 0. As the algorithm proceeds, observations are linked, represented by horizontal lines.

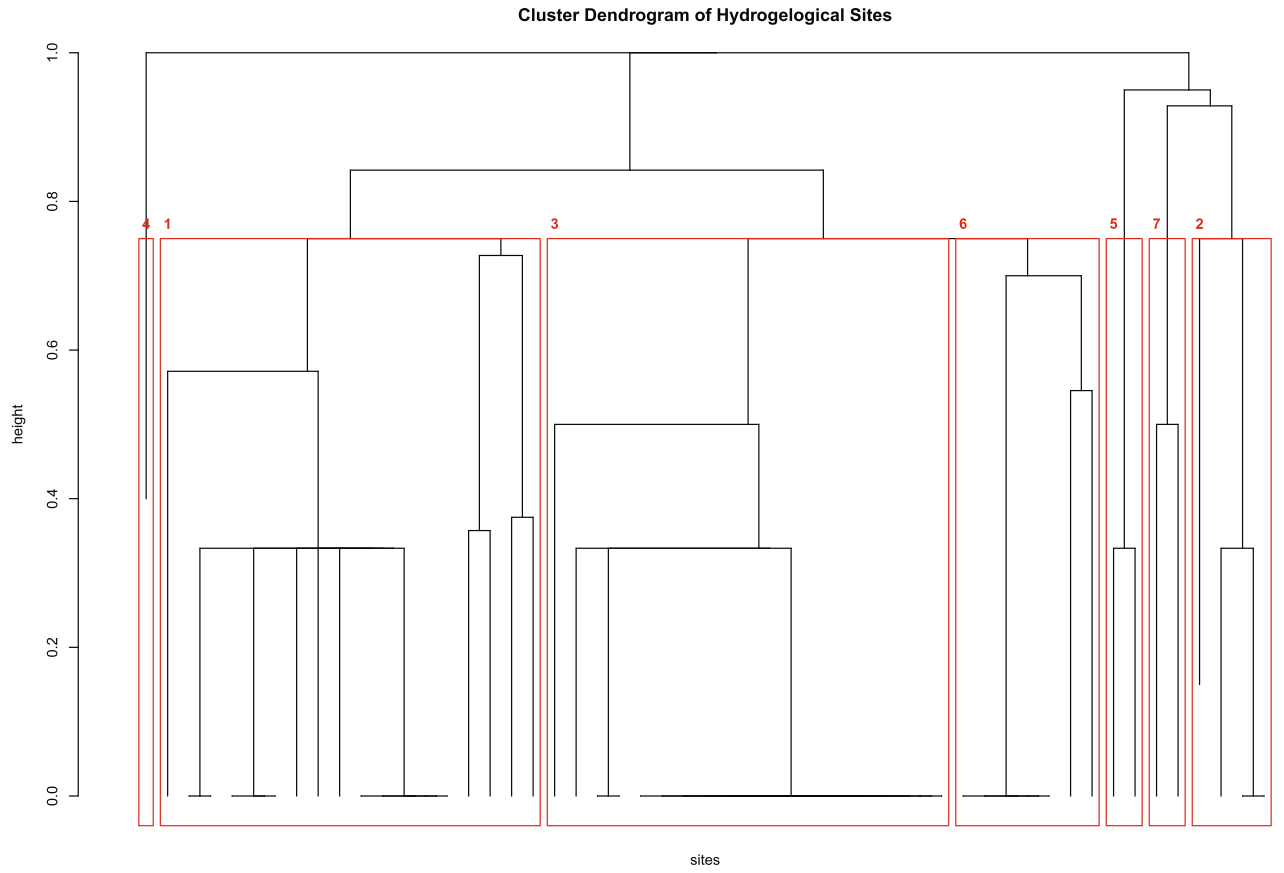

Figure 3. Hierarchical clustering dendrogram (black) of 53 hydrogeological sites. Each site begins at height 0 and is grouped with similar sites (a grouping is a horizontal bar). Cutting the dendrogram with a horizontal line at a given height yields a clustering. In this study, we cut the dendrogram into seven clusters. Each cluster is depicted by a red box.

Since several sites have a Jaccard distance of 0, they were linked successively to form one large cluster in the first few iterations of the algorithm. They form one large horizontal line at the very bottom of the dendrogram (see Figure 3). The complete linkage leads to the formation of round clusters, where groups were more likely to merge than singleton clusters. To select the k^* number of clusters, we computed $W_p(k)$ (see Equation 3) for k between 2 and 22. The k that minimized $W_p(k)$ were $k = 7$ and $k = 11$. We chose k^* to be the smaller of the two. Selecting a different number of clusters, as well as the possible impact of this choice, will be discussed below.

Table 1 shows the results of clustering, describing for each cluster its size, average within-cluster Jaccard distance, and silhouette coefficient. The silhouette coefficients of clusters 3, 5, and 6 are relatively large, meaning that these clusters are both highly similar within themselves and very different from the rest of the clusters. Clusters 1, 2, and 7 are less distinct from the rest of the sites, meaning that there are some sites in these clusters that could have been part of their neighbors. Cluster 4 has a silhouette width of 0 due to the fact that it has only one element.

Poor silhouette performance would indicate that a different number of clusters, or even a modification of the clustering method, would have produced better results. It is likely that a portion of the between-cluster similarity comes from the high frequency of a few attributes. The silhouettes reveal that the majority of observations belong to well-separated clusters, while a few observations could be placed into different clusters. However, as the final clustering was chosen with a specific application in mind, with emphasis on avoiding singleton clusters, we opt to leave the clustering as is.

Table 1
The Final Clustering of 53 Sites in the *WWHYPDA*

C	1	2	3	4	5	6	7
$\lvert C \rvert$	18	4	18	1	2	7	3
$s(C)$	0.27	0.33	0.73	0.00	0.64	0.46	0.36

Note: C denotes cluster (1 to 7), $\lvert C \rvert$ is the number of sites in each cluster, and $s(C)$ is the average silhouette width of cluster C.

Predictive Performance of Regionalized and Nonregionalized Priors: Using the clustering, we have now determined groups of similar sites. Next, we construct, for the 52 sites belonging to nonsingleton clusters, regionalized and nonregionalized priors. The nonregionalized priors are computed from all available measurements that do not belong to the target site.

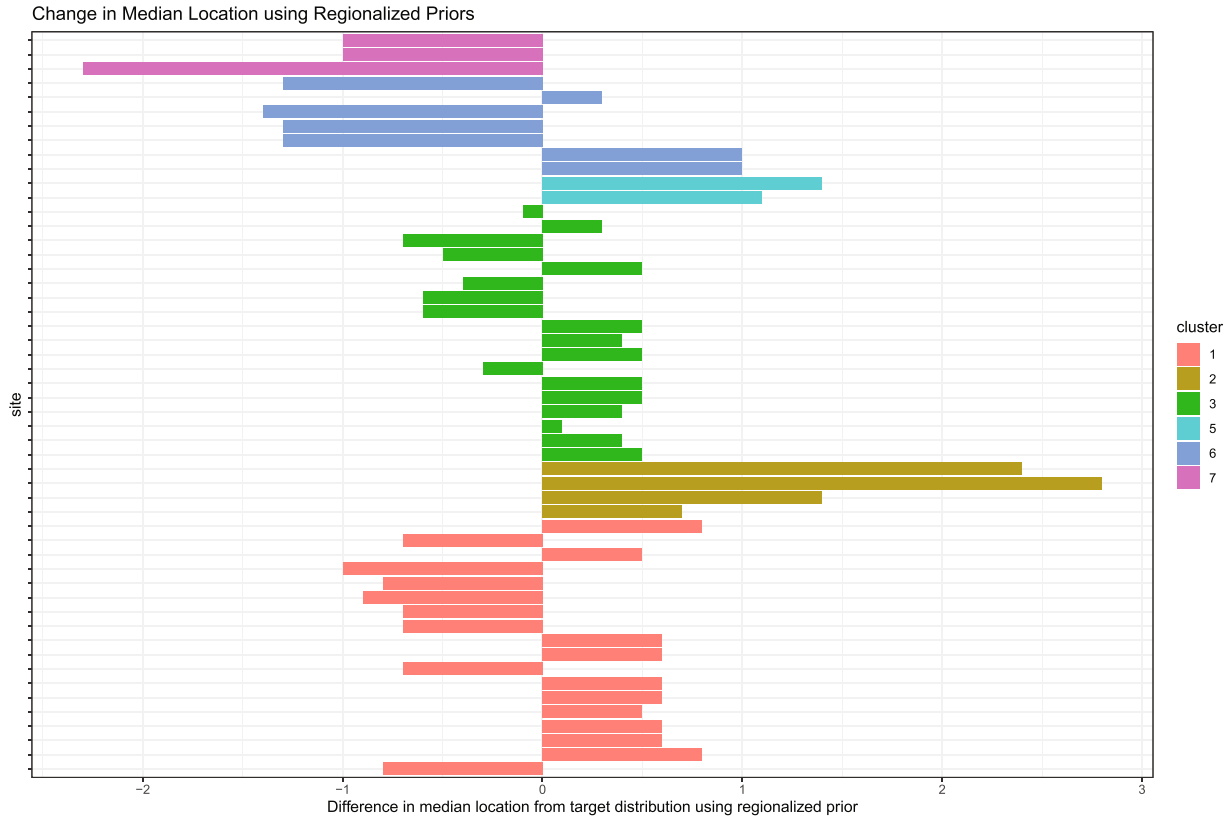

Figure 4. Evaluation of predictive performance of *exPrior* algorithm using regionalized priors, based on difference in location of median. Coloring represents the cluster the site belongs to.

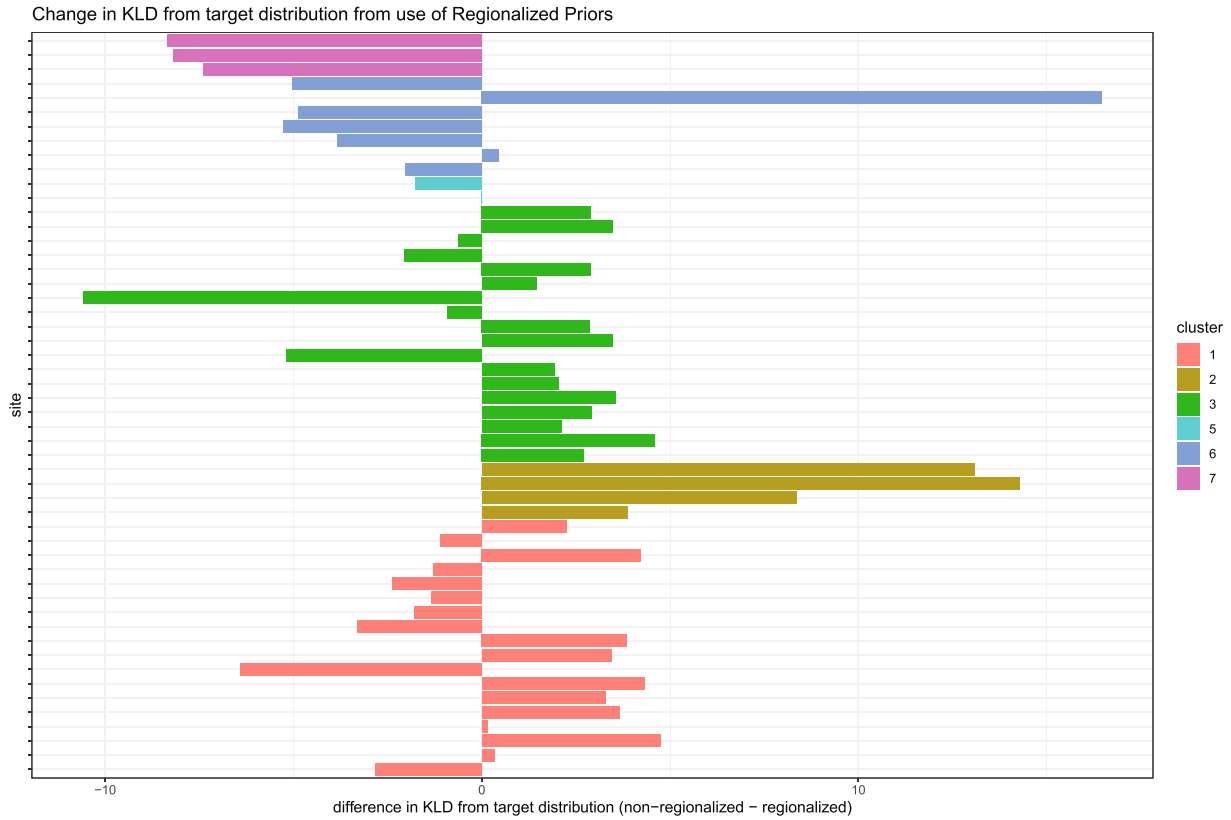

Figure 5. Evaluation of predictive performance of *exPrior* algorithm using regionalized priors, based on the difference in KLD from the target distribution (>0 is better, <0 is worse). Coloring represents the cluster the site belongs to.

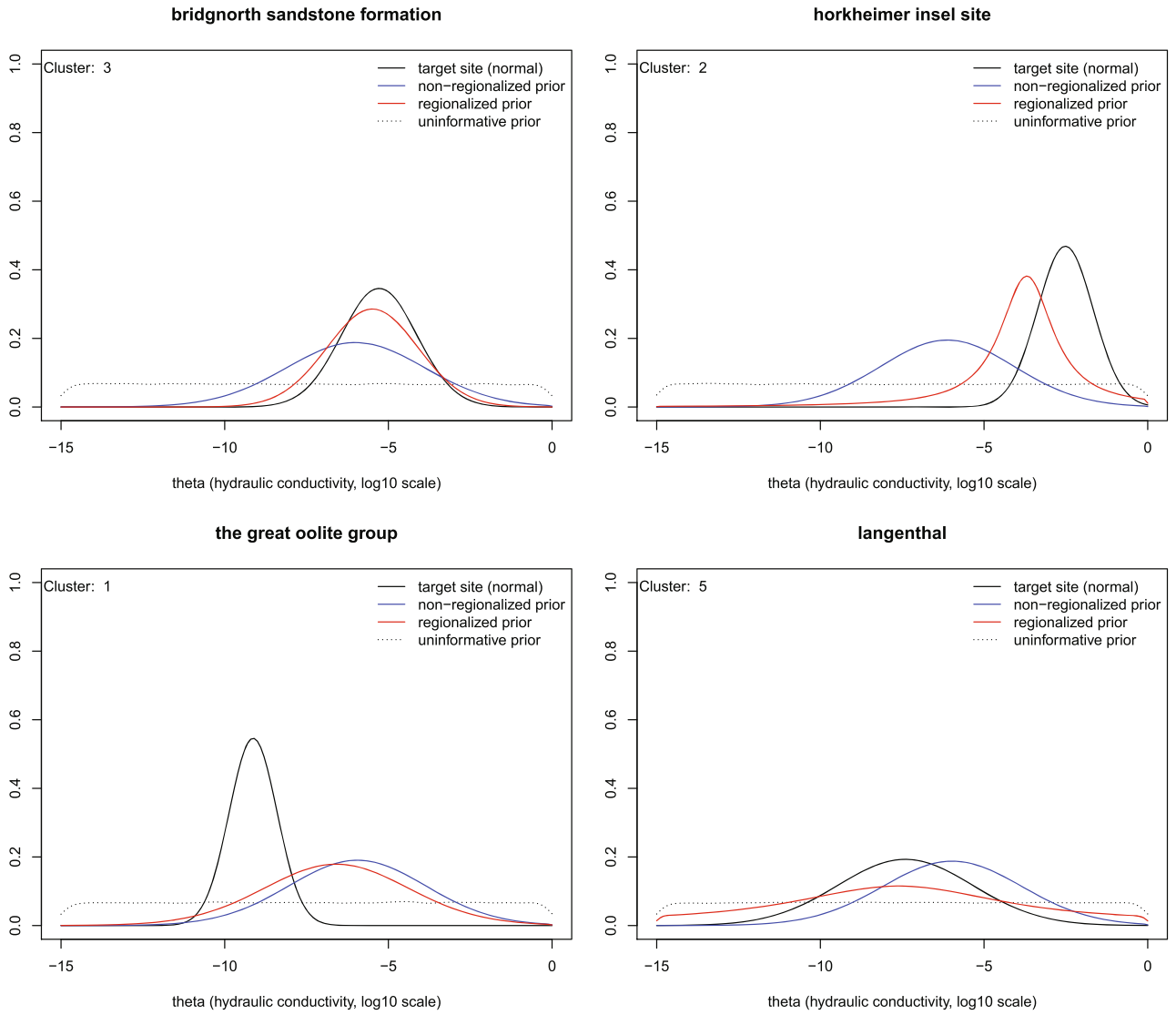

Figure 6. Comparison of the predictive performance of regionalized and nonregionalized priors for 4 of the 52 sites in our experiment. Starting on the top-left, we see an example from Cluster 3 where a regionalized prior is closer to the target distribution. Moving clockwise, we see the same trend, where a regionalized prior is more peaked and closer to the target. Continuing clockwise, we see examples where the regionalized prior is only a slight improvement on the nonregionalized prior.

The results are displayed in Figures 4 and 5, which display the predictive performance of the regionalized versus the nonregionalized priors for each site, using the difference in median location from the target site in Figure 4 and the difference in KLD from the target site in 5. The difference is computed by subtracting the quantity of interest median location (or KLD) computed with the regionalized prior from the one computed using the nonregionalized prior. In both figures, a positive value indicates an improvement (we are closer to the target distribution when using a regionalized prior), while a negative value indicates a loss in performance (we are farther from the target distribution when using a regionalized prior).

Example Cases: To visualize our results, we can plot for each target site its density (this could be a kernel density estimate, or a normal density with parameters obtain from the site's measurements) next to its derived regionalized and nonregionalized priors. As a basis of comparison, we also plot a flat prior. This allows us to see the impact of choosing a regionalized prior over a flat prior, as well as the difference between informative and noninformative priors. Figure 6 shows this for four selected sites.

As can be seen from our results, the regionalized priors typically outperformed their nonregionalized counterparts. However, not all sites showed a large improvement, if any at all (see examples in Figure 6). Here, the noninformative prior was arbitrary to some degree, but the distribution at the target site represents a measure of statistical uncertainty that cannot be overcome. Getting close to this distribution is therefore the best that could be achieved. In the examples, we can see a range of behavior. For some sites, the regionalized prior is very close to

the distribution of the target site, and a marked improvement on using a nonregionalized prior. In other sites, there is hardly any difference between regionalized and nonregionalized priors. In the former case, having more measurements is not necessary, as those from similar sites are sufficient in deriving an informative prior. In the latter case, having more descriptive information to characterize a site would improve the ability to select similar sites.

Discussion

The experimental results should be considered as a proof-of-concept for future large-scale hydrosystems analysis. In fact, with more data and domain knowledge incorporated into the clustering, there can be more information uncovered on the relationships between hydrogeological sites and the feasibility of data transfer. It is therefore important to asses areas for further consideration and experimentation.

We saw generally only modest improvements in the predictive accuracy of prior distributions derived with our method. We believe that this data-driven method could be more effective with an increase in the number of available hydrogeological measurements, as well as qualitative descriptions. This has ramification for the data collection efforts necessary for prior derivation in particular and data assimilation in general.

It is clear that the raw numbers of measurements that are available in a given dataset are an important criterion for the impact of any data-driven method. The features currently present in the *WWHYPDA* only represent a subset of attributes that characterize a site. For example, the used measurement technique or the fracture apertures are known to be important to characterizing a site, yet are still missing in the database. We expect that including such additional features will be as important as the inclusion of additional measurements for the improvement of data assimilation efforts. In general, the process of data generation, data collection and data assimilation should be a feedback loop. Our results therefore provide an important link in this loop by identifying gaps that currently exist and the lack of a broad range of features is currently such a gap. Since the method present here is easily reproducible by virtue of all scripts and tools to derive the results being openly available, it is feasible to assess the impact of adding features on the resulting clustering and derived prior distributions.

Additionally, the method can be adjusted to better reflect domain knowledge of site similarity. As aforementioned, both the distributions of the data and the distance metric used impact the resulting hierarchy of similar sites. In situations with a discrepancy in feature variance, one can apply transformations in order to reduce the possible impact of feature variance, especially when using nonbinary data and Euclidean distance to measure site distance. One can also incorporate expert information by weighting features based on their relevance to site similarity found in other experiments. Furthermore, a user can modify the method by taking into consideration different criteria for cutting an HAC dendrogram, such as setting a threshold for cluster size to avoid singleton clusters or considering a weighted measure of cluster distance. Such changes would potentially yield more accurate results and lead to a clustering that more closely reflects domain knowledge.

Our results help to delineate the most promising next steps in order to improve the predictive ability of Bayesian inference in hydrogeology. Given that lack of data was the single biggest challenge, it stands to reason that more effort needs to be put into changing that. After all, data are the building material of any kind of data-driven inference and the sophistication of the algorithm cannot overcome the limitations present in the used dataset (Halevy et al. 2009).

Conclusion

We introduce a method for using hierarchical clustering to define a notion of site similarity, demonstrating it sites with measurements and features present in the *WWHYPDA*. We apply our method to the derivation of ex situ priors introduced by Cucchi et al. (2019), using it as a data selection step. We conducted an experiment to test the efficacy of our method: that is, to see whether deriving ex situ priors from a small set of similar sites produces prior distributions that match those of target sites. We found mixed results; while ex situ priors are more informative than flat priors, deriving them using a set of similar sites was not always any better than deriving them with all available measurements. Possible reasons include a lack of available data to characterize sites, as well as room for improving the measurement of distance between sites. We conclude that there is now a unique opportunity to combine hydrogeological experience with data-driven methods applied to hydrogeological databases to develop data-driven measures of similarity and increase the data worth of readily-available hydrogeological information for future hydrogeological studies.

Author Contributions

Nura Kawa, Karina Cucchi and Falk Heße are the main authors of the `geostatDB` and `exPrior` package and are responsible for all parts of the manuscript. Nura Kawa is responsible for the analysis performed here. Yoram Rubin acted as the supervisor for Karina Cucchi and Nura Kawa and provided help for all portions of the manuscript. Sabine Attinger provided her assistance and expertise for the completion of the manuscript.

Acknowledgments

For this study, Falk Heße was financially supported by the Deutsche Forschungsgemeinschaft via Grant Number: HE-7028-1/2. Nura Kawa was funded by the Deutsche Forschungsgemeinschaft via Grant Number: HE-7028-1/2, by the Science@Leuven scholarship as well as by CRC 1076 AquaDiva. We are also thankful for

the contribution of the creators of the *WWHYPDA*. Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement

The data being used in this study were drawn from the WWHYPDA. We also used a number of software packages for the preparation, derivation and analysis of the results. They are as follows:

- To import the data into R, we used the R package `geostatDB` being developed at https://github.com/GeoStat-Bayesian/geostatDB.
- For the computation of the ex situ prior distribution, we used the R package `exPrior` being developed at https://github.com/GeoStat-Bayesian/exPrior. The used software version was 1.0.1 (Heße et al. 2021).
- The R code used for the analysis can be found at https://github.com/GeoStat-Bayesian/siteSimilarity.

Authors' Note

The authors do not have any conflicts of interest or financial disclosures to report.

References

Bandyopadhyay, S., and S. Saha. 2013. *Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications*, 1st ed. Berlin Heidelberg: Springer Link.

Blöschl, G., and M. Sivapalan. 1995. Scale issues in hydrological modelling: A review. *Hydrological Processes* 9, no. 3–4: 251–290.

Comunian, A., and P. Renard. 2009. Introducing wwhypda: A world-wide collaborative hydrogeological parameters database. *Hydrogeology Journal* 17, no. 2: 481–489.

Cucchi, K., F. Heße, N. Kawa, C. Wang, and Y. Rubin. 2019. Ex-situ priors: A Bayesian hierarchical framework for defining informative prior distributions in hydrogeology. *Advances in Water Resources* 126: 65–78.

Easwaran, K. 2011a. Bayesianism I: Introduction and arguments in favor. *Philosophy Compass* 6, no. 5: 312–320.

Easwaran, K. 2011b. Bayesianism II: Applications and criticisms. *Philosophy Compass* 6, no. 5: 321–332.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2014. *Bayesian Data Analysis. Statistical Science*, 2nd ed. Boca Raton, London, New York.: Chapman and Hall/CRC.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. 2013. *Bayesian data analysis. Statistical Science (Book 106)*, 3rd ed. Boca Raton, London, New York: Chapman and Hall/CRC.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1, no. 3: 515–534.

Hajek, A., and C. Hitchcock. 2016. *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

Hajek, A. 2007. The reference class problem is your problem too. *Synthese* 156, no. 3: 563–585 Workshop on Bayesian Epistemology, London Sch Econ & Polit Sci, Ctr Philosophy Nat & Social Sci, London, England.

Halevy, A., P. Norvig, and F. Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, no. 2: 8–12.

Halpern, J.Y. 2003. *Reasoning about Uncertainty*. Cambridge, London: MIT Press.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning. *Bayesian Forecasting and Dynamic Models* 1: 1–694.

Heße, F., K. Cucchi, N. Kawa, and Y. Rubin. 2021. exPrior: An R package for the formulation of ex-situ priors. *The R Journal* 13.

Heße, F., A. Comunian, and S. Attinger. 2019a. What we talk about when we talk about uncertainty. Toward a unified, data-driven framework for uncertainty characterization in hydrogeology. *Frontiers in Earth Science* 7.

Heße, F., K. Cucchi, and N. Kawa. 2019b. Geostat-bayesian/exprior: First release.

Heße, F., K. Cucchi, and N. Kawa. 2019c. Geostat-bayesian/geostatdb: First release.

Jaeger, M. 2006. A logic for inductive probabilistic reasoning. In *Uncertainty, Rationality, and Agency*, 11–79. Heidelberg Berlin: Springer.

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

Kumar, R., L. Samaniego, and S. Attinger. 2013. Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research* 49, no. 1: 360–379.

Li, X., Y. Li, C.-F. Chang, B. Tan, Z. Chen, J. Sege, C. Wang, and Y. Rubin. 2017. Stochastic, goal-oriented rapid impact modeling of uncertainty and environmental impacts in poorly-sampled sites using ex-situ priors. *Advances in Water Resources* 111, no. 1: 174–191.

Lung, C.H., and C. Zhou. 2010. Using hierarchical agglomerative clustering in wireless sensor networks: An energy-efficient and flexible approach. *Ad Hoc Networks* 8, no. 3: 328–344.

Manning, C.D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

McIntyre, N., H. Lee, H. Wheater, A. Young, and T. Wagener. 2005. Ensemble predictions of runoff in ungauged catchments. *Water Resources Research* 41, no. 12: W12434.

Merz, R., and G. Blöschl. 2005. Flood frequency regionalisation-spatial proximity vs. catchment attributes. *Journal of Hydrology* 302, no. 1–4: 283–306.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rousseeuw, P.J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20: 53–65.

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

Tang, Y., L. Marshall, A. Sharma, and T. Smith. 2016. Tools for investigating the prior distribution in Bayesian hydrology. *Journal of Hydrology* 538: 551–562.

Ulrych, T.J., M.D. Sacchi, and A. Woodbury. 2001. A Bayes tour of inversion: A tutorial. *Geophysics* 66, no. 1: 55–69.

Wallmann, C. 2017. A Bayesian solution to the conflict of narrowness and precision in direct inference. *Journal for General Philosophy of Science* 48: 485–500.

Zacharias, S., and G. Wessolek. 2007. Excluding organic matter content from pedotransfer predictors of soil water retention. *Soil Science Society of America Journal* 71, no. 1: 43–50.