



Geostatistical Learning: Challenges and Opportunities

Júlio Hoffmann^{*†}, Maciel Zortea, Breno de Carvalho and Bianca Zadrozny

IBM Research, Rio de Janeiro, Brazil

OPEN ACCESS

Edited by:

Dabao Zhang,
Purdue University, United States

Reviewed by:

Yunlong Feng,
University at Albany, United States
Mudasser Naseer,
University of Lahore, Pakistan

*Correspondence:

Júlio Hoffmann
julio.hoffmann@impa.br

†Present address:

Instituto de Matemática Pura e
Aplicada, Rio de Janeiro, Brazil

Specialty section:

This article was submitted to
Mathematics of Computation
and Data Science,
a section of the journal
Frontiers in Applied Mathematics and
Statistics

Received: 01 April 2021

Accepted: 04 June 2021

Published: 01 July 2021

Citation:

Hoffmann J, Zortea M, de Carvalho B
and Zadrozny B (2021) Geostatistical
Learning: Challenges
and Opportunities.
Front. Appl. Math. Stat. 7:689393.
doi: 10.3389/fams.2021.689393

Statistical learning theory provides the foundation to applied machine learning, and its various successful applications in computer vision, natural language processing and other scientific domains. The theory, however, does not take into account the unique challenges of performing statistical learning in geospatial settings. For instance, it is well known that model errors cannot be assumed to be independent and identically distributed in geospatial (a.k.a. regionalized) variables due to spatial correlation; and trends caused by geophysical processes lead to covariate shifts between the domain where the model was trained and the domain where it will be applied, which in turn harm the use of classical learning methodologies that rely on random samples of the data. In this work, we introduce the *geostatistical (transfer) learning* problem, and illustrate the challenges of learning from geospatial data by assessing widely-used methods for estimating generalization error of learning models, under covariate shift and spatial correlation. Experiments with synthetic Gaussian process data as well as with real data from geophysical surveys in New Zealand indicate that none of the methods are adequate for model selection in a geospatial context. We provide general guidelines regarding the choice of these methods in practice while new methods are being actively researched.

Keywords: geostatistical learning, transfer learning, covariate shift, geospatial, density ratio estimation, importance weighted cross-validation

1 INTRODUCTION

Classical learning theory [1–3] and its applied machine learning methods have been popularized in the geosciences after various technological advances, leading initiatives in open-source software [4–7], and intense marketing from a diverse portfolio of industries. In spite of its popularity, learning theory cannot be applied straightforwardly to solve problems in the geosciences as the characteristics of these problems violate fundamental assumptions used to derive the theory and related methods (e.g., i. i.d. samples).

Among these methods derived under classical assumptions (more on this later), those for estimating the generalization (or prediction) error of learned models in unseen samples are crucial in practice [2]. In fact, estimates of generalization error are widely used for selecting the best performing model for a problem, out of a collection of available models [8], and statistical learning can be posed broadly as minimization of generalization error. If estimates of error are inaccurate because of violated assumptions, then there is great chance that models will be selected inappropriately [9]. The issue is aggravated when models of great expressiveness (i.e., many learning parameters) are considered in the collection since they are quite capable of overfitting the available data [10, 11].

The literature on generalization error estimation methods is vast [8, 12], and we do not intend to review it extensively here. Nevertheless, some methods have gained popularity since their

introduction in the mid 70s because of their generality, ease of use, and availability in open-source software:

Leave-one-out (1974). The leave-one-out method for assessing and selecting learning models was based on the idea that to estimate the prediction error on an unseen sample one only needs to hide a seen sample from a dataset and learn the model. Because the hidden sample has a known label, the method can compare the model prediction with the true label for the sample. By repeating the process over the entire dataset, one gets an estimate of the expected generalization error [13]. Leave-one-out has been investigated in parallel by many statisticians, including Nicholson (1960) and Stone (1974), and is also known as ordinary cross-validation.

k-fold cross-validation (1975). The term k-fold cross-validation refers to a family of error estimation methods that split a dataset into non-overlapping “folds” for model evaluation. Similar to leave-one-out, each fold is hidden while the model is learned using the remaining folds. It can be thought of as a generalization of leave-one-out where folds may have more than a single sample [14, 15]. Cross-validation is less computationally expensive than leave-one-out depending on the size and number of folds, but can introduce bias in the error estimates if the number of samples in the folds used for learning is much smaller than the original number of samples in the dataset.

Major assumptions are involved in the derivation of the estimation methods listed above. The first of them is the assumption that samples come from independent and identically distributed (i.i.d.) random variables. It is well-known that spatial samples are not i.i.d., and that spatial correlation needs to be modeled explicitly with geostatistical theory. Even though the sample mean of the empirical error used in those methods is an unbiased estimator of the prediction error regardless of the i.i.d. assumption, the precision of the estimator can be degraded considerably with non-i.i.d. samples.

Motivated by the necessity to leverage non-i.i.d. samples in practical applications, and evidence that model’s performance is affected by spatial correlation [16, 17], the statistical community devised new error estimation methods using the spatial coordinates of the samples:

h-block leave-one-out (1995). Developed for time-series data (i.e., data showing temporal dependency), the h-block leave-one-out method is based on the principle that stationary processes achieve a correlation length (the “h”) after which the samples are not correlated. The time-series data is then split such that samples used for error evaluation are at least “h steps” distant from the samples used to learn the model [18]. Burman (1994) showed how the method outperformed traditional leave-one-out in time-series prediction by selecting the hyperparameter “h” as a fraction of the data, and correcting the error estimates accordingly to avoid bias.

Spatial leave-one-out (2014). Spatial leave-one-out is a generalization of h-block leave-one-out from time-series to spatial data [19]. The principle is the same, except that the blocks have multiple dimensions (e.g., norm-balls).

Block cross-validation (2016). Similarly to k-fold cross-validation for non-spatial data, block cross-validation was proposed as a faster alternative to spatial leave-one-out. The

method creates folds using blocks of size equal to the spatial correlation length, and separates samples for error evaluation from samples used to learn the model. The method introduces the concept of “dead zones”, which are regions near the evaluation block that are discarded to avoid over-optimistic error estimates [20, 21].

Unlike the estimation methods proposed in the 70s, which use random splits of the data, these methods split the data based on spatial coordinates and what the authors called “dead zones”. This set of heuristics for creating data splits avoids configurations in which the model is evaluated on samples that are too near (< spatial correlation length) other samples used for learning the model. Consequently, these estimation methods tend to produce error estimates that are higher on average than their non-spatial counterparts, which are known to be over-optimistic in the presence of spatial correlation. However, systematic splits of the data introduce bias, which have not been emphasized enough in the literature.

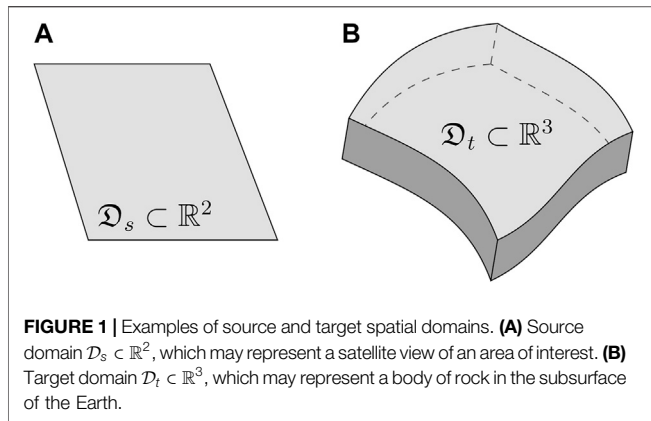
All methods for estimating generalization error in classical learning theory, including the methods listed above, rely on a second major assumption. The assumption that the distribution of unseen samples to which the model will be applied is equal to the distribution of samples over which the model was trained. This assumption is very unrealistic for various applications in geosciences, which involve quite heterogeneous (i.e., variable), heteroscedastic (i.e., with different variability) processes [22].

Very recently, an alternative to classical learning theory has been proposed, known as transfer learning theory, to deal with the more difficult problem of learning under shifts in distribution, and learning tasks [23–25]. The theory introduces methods that are more amenable for geoscientific work [26–28], yet these same methods were not derived for geospatial data (e.g. climate data, Earth observation data, field measurements).

Of particular interest in this work, the covariate shift problem is a type of transfer learning problem where the samples on which the model is applied have a distribution of covariates that differs from the distribution of covariates over which the model was trained [29]. It is relevant in geoscientific applications in which a list of explanatory features is known to predict a response via a set of physical laws that hold everywhere. Under covariate shift, a generalization error estimation method has been proposed:

Importance-weighted cross-validation (2007). Under covariate shift, and assuming that learning models may be misspecified, classical cross-validation is not unbiased. Importance weights can be considered for each sample to recover the unbiasedness property of the method, and this is the core idea of importance-weighted cross-validation [30, 31]. The method is unbiased under covariate shift for the two most common supervised learning tasks: regression and classification.

The importance weights used in importance-weighted cross-validation are ratios between the target (or test) probability density and the source (or train) probability density of covariates. Density ratios are useful in a much broader set of applications including two-sample tests, outlier detection, and distribution comparison. For that reason, the problem of density ratio estimation has become a general statistical problem [32]. Various density ratio estimators have been proposed with



increasing performance [33–36], yet an investigation is missing that contemplates importance-weighted cross-validation and other existing error estimation methods in geospatial settings.

In this work, we introduce *geostatistical (transfer) learning*, and discuss how most prior work in spatial statistics fits in a specific type of learning from geospatial data that we term *pointwise learning*. In order to illustrate the challenges of learning from geospatial data, we assess existing estimators of generalization error from the literature using synthetic Gaussian process data and real data from geophysical well logs in New Zealand that we made publicly available [37].

The paper is organized as follows. In **Section 2**, we introduce *geostatistical (transfer) learning*, which contains all the elements involved in learning from geospatial data. We define covariate shift in the geospatial setting and briefly review the concept of spatial correlation. In **Section 3**, we define generalization error in geostatistical learning, discuss how it generalizes the classical definition of error in non-spatial settings, and review estimators of generalization error from the literature devised for *pointwise learning*. In **Section 4**, we present our experiments with geospatial data, and discuss the results of different error estimation methods. In **Section 5**, we conclude the work and share a few remarks regarding the choice of error estimation methods in practice.

2 GEOSTATISTICAL LEARNING

In this section, we define the elements of statistical learning in geospatial settings. We discuss the covariate shift and spatial correlation properties of the problem, and illustrate how they affect the involved feature spaces.

Consider a sample space Ω , a source spatial domain $\mathcal{D}_s \subset \mathbb{R}^{d_s}$, and a target spatial domain $\mathcal{D}_t \subset \mathbb{R}^{d_t}$ on which stochastic processes (i.e., spatial random variables) are defined:

$$\begin{aligned} Z_{s_j} : \mathcal{D}_s \times \Omega &\rightarrow \mathbb{R}, j = 1, 2, \dots, n_s \text{ on source domain } \mathcal{D}_s \\ Z_{t_j} : \mathcal{D}_t \times \Omega &\rightarrow \mathbb{R}, j = 1, 2, \dots, n_t \text{ on target domain } \mathcal{D}_t \end{aligned} \quad (1)$$

For example, $(Z_{s_j})_{j=1,2,\dots,n_s}$ may represent a collection of processes observed remotely from satellite on a 2D surface

$\mathcal{D}_s \subset \mathbb{R}^2$, whereas $(Z_{t_j})_{j=1,2,\dots,n_t}$ may represent processes that occur within the 3D subsurface of the Earth $\mathcal{D}_t \subset \mathbb{R}^3$ (see **Figure 1**). Any process Z in these collections can be viewed in two distinct ways:

Geostatistical theory. From the viewpoint of geostatistical theory, samples $z(\cdot, \omega)$ of the process $Z(\mathbf{u}, \omega)$ are obtained by fixing $\omega \in \Omega$. These samples are spatial maps that assign a real number to each location $\mathbf{u} \in \mathcal{D}$.

Learning theory. From the viewpoint of statistical learning theory, scalar samples $z(\mathbf{u}, \cdot)$ are obtained by fixing $\mathbf{u} \in \mathcal{D}$. These scalar samples are ordered into a “feature vector” $\mathbf{x}_{\mathbf{u}} = (z_1, z_2, \dots, z_n)$ for a collection of processes $(Z_j)_{j=1,2,\dots,n}$ and for a specific location $\mathbf{u} \in \mathcal{D}$. In this case, $\mathbf{X}_{\mathbf{u}} : \Omega \rightarrow \mathbb{R}^n$ denotes the corresponding random vector of features such that $\mathbf{x}_{\mathbf{u}} \sim \mathbf{X}_{\mathbf{u}}$.

In order to define the geostatistical learning problem, we need to understand the joint probability distribution of features for all locations in a spatial domain $\Pr(\{\mathbf{X}_{\mathbf{u}}\}_{\mathbf{u} \in \mathcal{D}})$. This distribution is very complex in general as feature vectors $\mathbf{X}_{\mathbf{u}}$ and $\mathbf{X}_{\mathbf{v}}$ for two different locations $\mathbf{u} \neq \mathbf{v}$ are not independent. The closer the locations $\mathbf{u}, \mathbf{v} \in \mathcal{D}$ in the spatial domain, the more similar are their features $\mathbf{x}_{\mathbf{u}}, \mathbf{x}_{\mathbf{v}} \in \mathbb{R}^n$ in the feature space. Moreover, given that only one realization $z^{obs} = z(\cdot, \omega) \sim Z$ of the process is available at any given time, one must introduce stationarity assumptions inside \mathcal{D} to pool together different scalar samples $z(\mathbf{u}, \cdot)$ from different locations $\mathbf{u} \in \mathcal{D}$ in the spatial domain, and be able to estimate the distribution.

Regardless of the stationarity assumptions involved in modeling these processes, we can assume that inside \mathcal{D} the probability $\Pr_{\mathcal{D}}(\mathbf{X}) = \Pr(\{\mathbf{X}_{\mathbf{u}}\}_{\mathbf{u} \in \mathcal{D}})$ is well-defined. For example, most prior art in statistical learning with geospatial data assume that the pointwise probability of features $\Pr_{\mathbf{u}}(\mathbf{X}) = \Pr(\mathbf{X}_{\mathbf{u}})$ is not a function of location, that is $\Pr_{\mathbf{u}}(\mathbf{X}) = \Pr(\mathbf{X}), \forall \mathbf{u} \in \mathcal{D}$. Under this assumption, samples from everywhere in \mathcal{D} are used to estimate $\Pr(\mathbf{X}) = \Pr(Z_1, Z_2, \dots, Z_n)$. With the additional assumption that the feature vectors $\mathbf{X}_{\mathbf{u}}$ and $\mathbf{X}_{\mathbf{v}}$ are independent, the joint distribution of features for all locations can be written as $\Pr_{\mathcal{D}}(\mathbf{X}) = \prod_{\mathbf{u} \in \mathcal{D}} \Pr_{\mathbf{u}}(\mathbf{X})$.

Whereas the pointwise stationarity assumption may be reasonable inside a given spatial domain, the assumption of spatial independence of features is rarely defensible in practice. Additionally, pointwise stationarity often does not transfer from a source domain \mathcal{D}_s where the model is learned to a target domain \mathcal{D}_t where the model is applied, and consequently the joint distributions of features differ $\Pr_{\mathcal{D}_s} \neq \Pr_{\mathcal{D}_t}$. Before we can illustrate these two issues in more detail, we need to complete the definition of geostatistical learning problems by introducing the notion of spatial learning tasks.

We have introduced the notion of spatial domain \mathcal{D} , and the notion of joint probability of features $\Pr_{\mathcal{D}}(\mathbf{X})$ for all locations in the domain. Now we introduce the notion of spatial learning tasks, which are similar to classical learning tasks, but with the main difference that they can leverage properties of the underlying spatial domain. Classically, a learning task describes an action in terms of available features to produce new data. For example, “predict feature Z_{j_0} from features (Z_{j_1}, Z_{j_2}) ,” or “cluster the samples using features $(Z_{j_1}, Z_{j_2}, Z_{j_3})$ ” are classical learning tasks. Differently,

a spatial learning task T involves the spatial domain \mathcal{D} besides the features, and is therefore more complex. Practical examples from the industry include:

- Mining: The task of segmenting a mineral deposit from drillhole samples using a set of features is a spatial learning task. It assumes the segmentation result to be a *contiguous volume* of rock, which is an additional constraint in terms of spatial coordinates.
- Agriculture: The task of identifying crops from satellite images is a spatial learning task. Locations that have the same crop type *appear together* in the images despite possible noise in image layers (e.g. presence of clouds, animals).
- Petroleum: The task of segmenting formations from seismic data is a spatial learning task because these formations are large-scale *near-horizontal* layers of stacked rock.

Many more examples of spatial learning tasks exist, and others are yet to be proposed. Given the concepts introduced above, we are now ready for the main definition of this section:

Definition (Geostatistical Learning). Let \mathcal{D}_s be a source spatial domain, and \mathcal{D}_t be a target spatial domain. Let $\text{Pr}_{\mathcal{D}_s}(\mathbf{X}_s)$ and $\text{Pr}_{\mathcal{D}_t}(\mathbf{X}_t)$ be the joint distributions of features for all locations in these domains, and let T_s and T_t be two spatial learning tasks. Geostatistical (transfer) learning consists of learning T_t over \mathcal{D}_t using the knowledge acquired while learning T_s over \mathcal{D}_s , and assuming that the observed spatial data in \mathcal{D}_s and \mathcal{D}_t are both a single spatial sample of $\text{Pr}_{\mathcal{D}_s}(\mathbf{X}_s)$ and $\text{Pr}_{\mathcal{D}_t}(\mathbf{X}_t)$, respectively.

There are considerable differences between the classical definition of transfer learning [23, 24], and the proposed definition above. First, the distribution we have denoted by $\text{Pr}_{\mathcal{D}}(\mathbf{X})$ is spatial and involves all the locations $\mathbf{u} \in \mathcal{D}$, whereas the distribution in classical transfer learning is the marginal for any specific location, obtained from the assumption of pointwise stationarity $\text{Pr}(\mathbf{X}_{\mathbf{u}}) = \text{Pr}(\mathbf{X})$. Second, we use the term domain to refer to spatial domains \mathcal{D} , whereas the non-spatial literature uses the same term for the pair $(\mathbf{X}_{\mathbf{u}}, \text{Pr}(\mathbf{X}_{\mathbf{u}})) = (\mathbf{X}, \text{Pr}(\mathbf{X}))$. Third, the spatial learning task T we have introduced may be described in terms of properties of the spatial domain, which are not available in generic transfer learning problems.

Having understood the main differences between classical and geostatistical learning, we now focus our attention to a specific type of geostatistical transfer learning problem, and illustrate some of the unique challenges caused by spatial dependence.

2.1 Covariate Shift

Assume that the two spatial domains are different $\mathcal{D}_s \neq \mathcal{D}_t$, but that they share a set of processes (Z_1, Z_2, \dots, Z_n) . Additionally, assume that pointwise stationarity holds. Let $Z_o = f(Z_1, Z_2, \dots, Z_n)$ be a new process obtained as a function of the shared processes, and assume that it has only been observed in \mathcal{D}_s via a measuring device and/or manual labeling. That is, $z_o^{obs}(\cdot, \omega) \sim Z_o$ is a spatial sample of the process Z_o over \mathcal{D}_s . Under these assumptions, we can introduce the shared vector of features

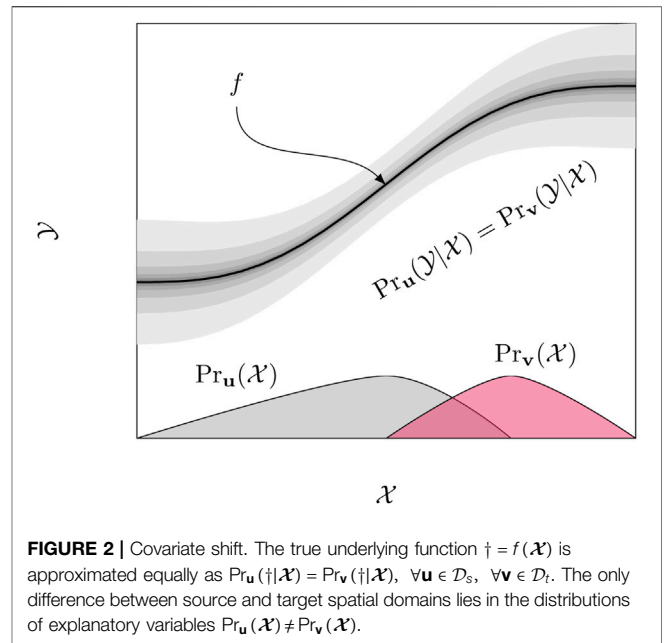


FIGURE 2 | Covariate shift. The true underlying function $\dagger = f(\mathcal{X})$ is approximated equally as $\text{Pr}_u(\dagger|\mathcal{X}) = \text{Pr}_v(\dagger|\mathcal{X})$, $\forall \mathbf{u} \in \mathcal{D}_s$, $\forall \mathbf{v} \in \mathcal{D}_t$. The only difference between source and target spatial domains lies in the distributions of explanatory variables $\text{Pr}_u(\mathcal{X}) \neq \text{Pr}_v(\mathcal{X})$.

$\mathbf{X}_s = \mathbf{X}_t = \mathbf{X} = (Z_1, Z_2, \dots, Z_n, Z_o)$, and the supervised learning task $T_s = T_t = T$ of predicting the process Z_o regardless of location $\mathbf{u} \in \mathcal{D}_s \cup \mathcal{D}_t$.

Let $\mathcal{X} = \mathbf{X}_{1:n}$ be the explanatory features, and $\mathcal{Y} = \mathbf{X}_{n+1}$ be the response feature. For any $\mathbf{u} \in \mathcal{D}_s$, we can write $\text{Pr}_u(\mathcal{X}, \mathcal{Y}) = \text{Pr}_u(\mathcal{Y}|\mathcal{X})\text{Pr}_u(\mathcal{X})$. Likewise, for any $\mathbf{v} \in \mathcal{D}_t$ we can write $\text{Pr}_v(\mathcal{X}, \mathcal{Y}) = \text{Pr}_v(\mathcal{Y}|\mathcal{X})\text{Pr}_v(\mathcal{X})$. The covariate shift property is defined as follows:

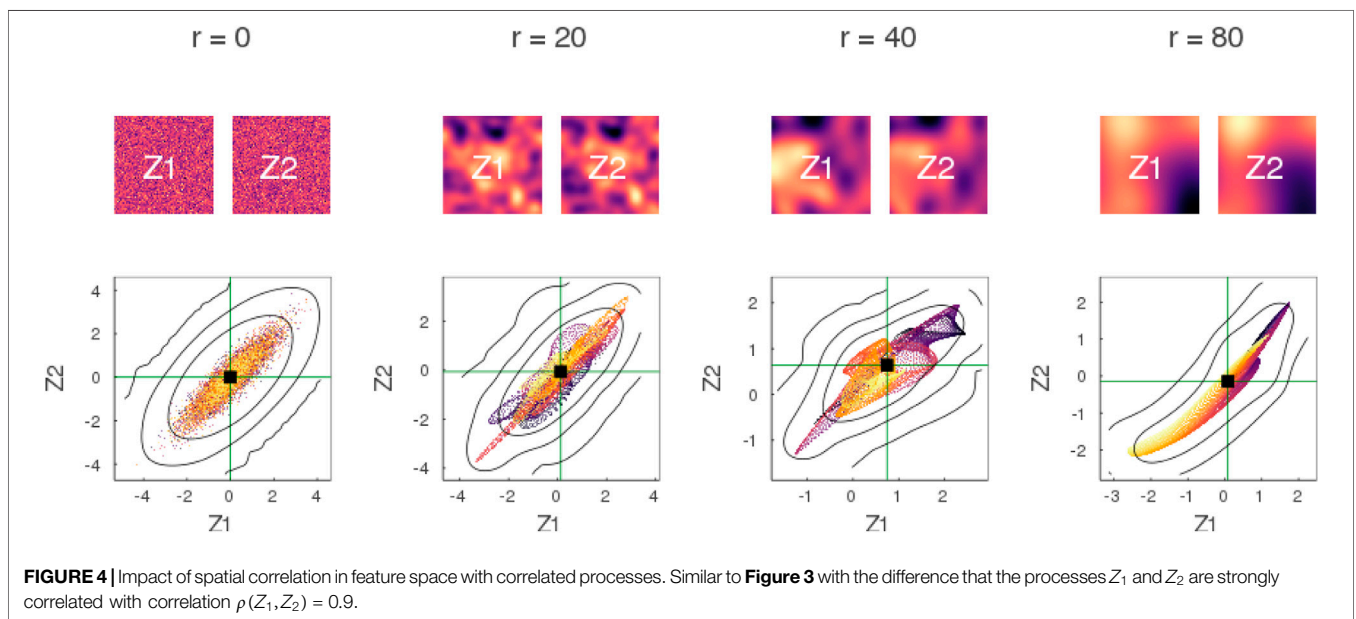
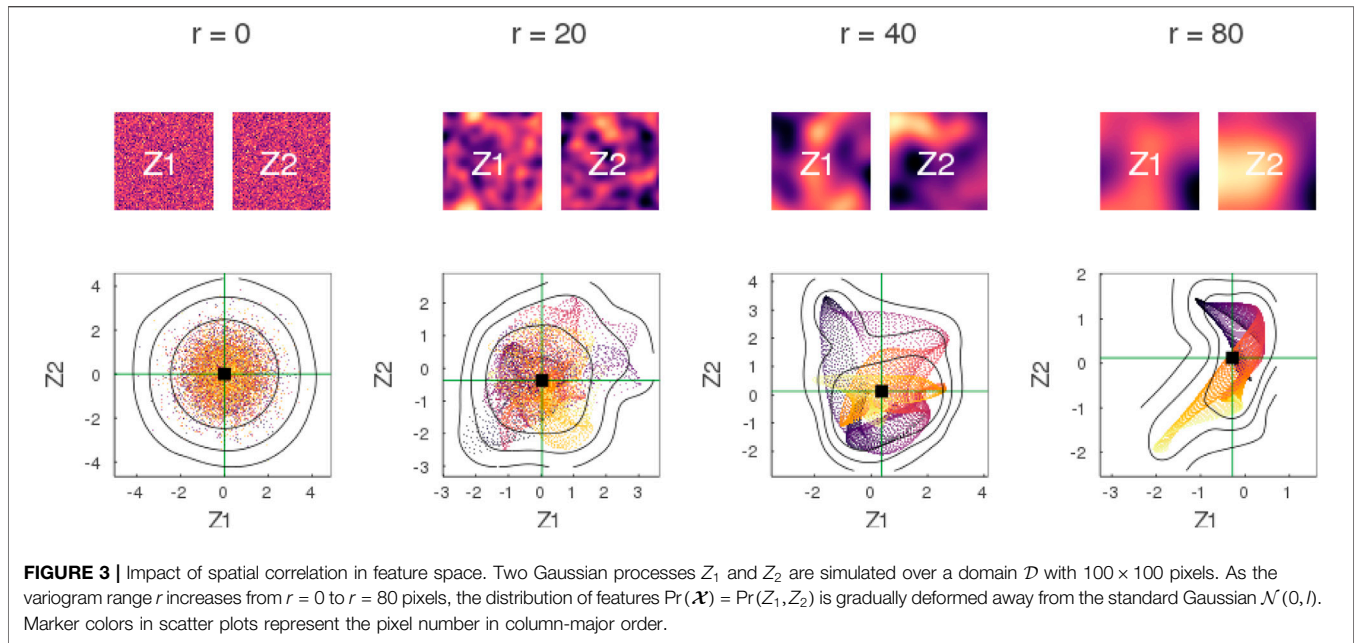
Definition (Covariate Shift). A geostatistical learning problem has the covariate shift property when for any $\mathbf{u} \in \mathcal{D}_s$ and for any $\mathbf{v} \in \mathcal{D}_t$ the distributions $\text{Pr}_u(\mathcal{X}, \mathcal{Y})$ and $\text{Pr}_v(\mathcal{X}, \mathcal{Y})$ differ by $\text{Pr}_u(\mathcal{X}) \neq \text{Pr}_v(\mathcal{X})$ while $\text{Pr}_u(\mathcal{Y}|\mathcal{X}) = \text{Pr}_v(\mathcal{Y}|\mathcal{X})$.

The property is based on the idea that the underlying true function f that created the process $\mathcal{Y} = f(\mathcal{X})$ is the same for all $\mathbf{u} \in \mathcal{D}_s$ and all $\mathbf{v} \in \mathcal{D}_t$. In this case, the function is approximated by the conditional distribution $\text{Pr}_u(\mathcal{Y}|\mathcal{X}) = \text{Pr}_v(\mathcal{Y}|\mathcal{X})$ for each and every location (see Figure 2).

In the geosciences, it is very common to encounter problems with covariate shift due to the great variability of natural processes. Whenever a model is 1) learned using labels provided by experts on a spatial domain \mathcal{D}_s , is 2) validated with classical train-validation-test methodologies (meaning that it satisfies some performance threshold), and yet 3) performs poorly on a new spatial domain \mathcal{D}_t where the labeling function is expected to be the same, we can conclude that there are shifts in distribution. In Section 4 we illustrate covariate shifts in real data that we prepared in-house from geophysical surveys in New Zealand.

2.2 Spatial Correlation

Another important issue with geospatial data that is often ignored is spatial dependence, which we illustrate next. As mentioned earlier, the closer are two locations $\mathbf{u}, \mathbf{v} \in \mathcal{D}$ in a spatial domain, the more similar are their features $\mathbf{x}_{\mathbf{u}}, \mathbf{x}_{\mathbf{v}} \in \mathbb{R}^n$ in the feature space. Different statistics are available to quantify this spatial



dependence in a collection of samples, and a popular choice from geostatistics is the variogram $\gamma(h)$, which estimates for each spatial lag $h = \|\mathbf{u} - \mathbf{v}\| \in \mathbb{R}_0^+$ a correlation $\sigma^2 - \gamma(h)$, where σ^2 is the total sill in the samples [22]. Parallel algorithms for efficient variogram estimation exist in the literature [38], and can be useful tools for fast diagnosis of the spatial correlation property:

Definition (Spatial Correlation). A geostatistical learning problem has the spatial correlation property when the variogram of any of the stochastic processes $(Z_{s_j})_{j=1,2,\dots,n_s}$ and $(Z_{t_j})_{j=1,2,\dots,n_t}$ defined over \mathcal{D}_s and \mathcal{D}_t has a non-negligible positive range (or correlation length)

Besides serving as a tool for diagnosing spatial correlation in geostatistical learning problems, variograms can also be used to

simulate spatial processes with theoretical correlation structure. In **Figure 3**, we illustrate the impact of spatial correlation in the feature space of two independent spatial processes Z_1 and Z_2 simulated with direct (a.k.a. LU) Gaussian simulation [39]. As we increase the variogram range r in a spatial domain \mathcal{D} with 100×100 pixels, we observe that the distribution of features $\Pr(\mathcal{X}) = \Pr(Z_1, Z_2)$ is gradually deformed from a standard Gaussian ($r = 0$) to a “boomerang”-shaped distribution ($r = 80$).

Similar deformations are observed when the two processes Z_1 and Z_2 are correlated. In **Figure 4**, we illustrate the impact of spatial correlation for an inter-process correlation of $\rho(Z_1, Z_2) = 0.9$.

Spatial correlation may have different impact in source and target domains \mathcal{D}_s and \mathcal{D}_t , and can certainly affect the generalization error of learning models. In our experiments of **Section 4**, we assume that the variogram range of source and target processes are equal (i.e. $r_s = r_t = r$) to facilitate the analysis of the results. In practice, source and target processes may also have different spatial correlation, which is a type of shift that is not considered in classical transfer learning problems.

3 GENERALIZATION ERROR OF LEARNING MODELS

Having defined geostatistical learning problems, and their covariate shift and spatial correlation properties, we now turn into a general definition of generalization error of learning models in geospatial settings. We review an importance-weighted approximation of a related generalization error based on pointwise stationarity assumptions, and the use of an efficient importance-weighted cross-validation method for error estimation.

Consider a geostatistical learning problem $\mathcal{P} = \{(\mathcal{D}_s, \text{Pr}_{\mathcal{D}_s}, T_s), (\mathcal{D}_t, \text{Pr}_{\mathcal{D}_t}, T_t)\}$ with a single supervised spatial learning task $T_s = T_t = T$ (e.g. regression), and assume that a set of response features \mathcal{Y}_u are created by a function f , based on a set of explanatory features \mathcal{X}_u for each and every location $u \in \mathcal{D}_s \cup \mathcal{D}_t$. Our goal is to learn a model $\{\mathcal{Y}_u\}_{u \in \mathcal{D}_t} \approx \hat{f}(\{\mathcal{X}_u\}_{u \in \mathcal{D}_t})$ over the target domain \mathcal{D}_t that approximates f in terms of expected risk for some spatial supervised loss function \mathcal{L} :

$$\hat{f} = \arg \min_g \mathbb{E}_{\text{Pr}_{\mathcal{D}_t}} [\mathcal{L}(\{\mathcal{Y}_u\}_{u \in \mathcal{D}_t}, g(\{\mathcal{X}_u\}_{u \in \mathcal{D}_t}))] \quad (2)$$

In the expected value of **Eq 2**, spatial samples of the processes are drawn from $\text{Pr}_{\mathcal{D}_t}$ and rearranged into feature vectors \mathcal{X}_u and \mathcal{Y}_u for every location $u \in \mathcal{D}_t$. The spatial loss function \mathcal{L} compares the spatial map of features from the sample $\{\mathcal{Y}_u\}_{u \in \mathcal{D}_t}$ with the approximated map from the model $g(\{\mathcal{X}_u\}_{u \in \mathcal{D}_t})$. The model \hat{f} is the model that minimizes the expected loss (or risk) over the target domain \mathcal{D}_t .

Definition (Generalization Error). The generalization error of a learning model \hat{f} in a geostatistical learning problem \mathcal{P} is the expected risk attained by the model when spatial samples are drawn from $\text{Pr}_{\mathcal{D}_t}$ over the target domain \mathcal{D}_t (see **Eq 2**).

Unlike the classical definition of generalization error, the definition above for geostatistical learning problems relies on a spatial loss function \mathcal{L} , and on spatial samples like those produced via geostatistical simulation [40, 41]. For truly spatial learning models \hat{f} that use multiple locations in the spatial domain to make predictions, this generalization error is more appropriate. In this present work, however; we do not target spatial learning models, and only consider pointwise learning:

Definition (Pointwise Learning). Given a family of classical (non-spatial) learning models $\{\hat{f}_u\}_{u \in \mathcal{D}}$, pointwise learning consists of learning the model $\hat{f}(\{\mathcal{X}_u\}_{u \in \mathcal{D}}) = \{\hat{f}_u(\mathcal{X}_u)\}_{u \in \mathcal{D}}$ that assigns for each location $u \in \mathcal{D}$ the value $\hat{f}_u(\mathcal{X}_u)$ independently of the explanatory features at other locations.

More specifically, we consider pointwise learning with families that are made of a single learning model $\{\hat{f}_u\}_{u \in \mathcal{D}} = \{f\}$. In this case,

the model f is often learned based on pointwise stationarity assumptions, for some pointwise loss $\hat{\mathcal{L}}$:

$$\hat{f} = \arg \min_g \mathbb{E}_{\text{Pr}} [\hat{\mathcal{L}}(\mathcal{Y}, g(\mathcal{X}))] \quad (3)$$

Although pointwise learning with a single model is a very simple type of geostatistical learning, it is by far the most widely used approach in the geospatial literature. We acknowledge this fact, and consider an empirical approximation of the pointwise expected risk in **Eq 3** as opposed to the spatial expected risk in **Eq 2**.

An empirical approximation of the pointwise expected risk of a model g can be obtained via discretization of the target spatial domain \mathcal{D}_t :

$$\mathcal{R}_t(g) = \mathbb{E}_{\text{Pr}_{u \in \mathcal{D}_t}} [\hat{\mathcal{L}}(\mathcal{Y}, g(\mathcal{X}))] \approx \frac{1}{|\mathcal{D}_t|} \sum_{u \in \mathcal{D}_t} \hat{\mathcal{L}}(\mathcal{Y}_u, g(\mathcal{X}_u)) \quad (4)$$

with $|\mathcal{D}_t|$ the number of locations in the discretization. The problem with this empirical approximation is that the response features \mathcal{Y}_u are not available in the target domain where the model will be applied. However, it is easy to show that the pointwise expected risk in **Eq 4** can be rewritten with importance weights $\dot{w}(\mathcal{X}, \mathcal{Y})$ when samples from \mathcal{D}_s are drawn instead [30]:

$$\mathcal{R}_t(g) = \mathbb{E}_{\text{Pr}_{u \in \mathcal{D}_t}} [\hat{\mathcal{L}}(\mathcal{Y}, g(\mathcal{X}))] = \mathbb{E}_{\text{Pr}_{u \in \mathcal{D}_s}} [\dot{w}(\mathcal{X}, \mathcal{Y}) \hat{\mathcal{L}}(\mathcal{Y}, g(\mathcal{X}))] \quad (5)$$

with $\dot{w}(\mathcal{X}, \mathcal{Y}) = \frac{\text{Pr}_{u \in \mathcal{D}_t}(\mathcal{X}, \mathcal{Y})}{\text{Pr}_{u \in \mathcal{D}_s}(\mathcal{X}, \mathcal{Y})}$. Under covariate shift, the importance weights only depend on the distribution of explanatory features $\dot{w}(\mathcal{X}) = \frac{\text{Pr}_{u \in \mathcal{D}_t}(\mathcal{X})}{\text{Pr}_{u \in \mathcal{D}_s}(\mathcal{X})}$, and we can write a simple importance-weighted empirical approximation:

$$\begin{aligned} \mathcal{R}_t(g) &= \mathbb{E}_{\text{Pr}_{u \in \mathcal{D}_s}} [\dot{w}(\mathcal{X}) \hat{\mathcal{L}}(\mathcal{Y}, g(\mathcal{X}))] \\ &\approx \frac{1}{|\mathcal{D}_s|} \sum_{u \in \mathcal{D}_s} \dot{w}(\mathcal{X}_u) \hat{\mathcal{L}}(\mathcal{Y}_u, g(\mathcal{X}_u)) \end{aligned} \quad (6)$$

Our goal is to find the pointwise model that minimizes the empirical risk approximation $\hat{\mathcal{R}}_t(g)$ introduced in **Eq 6**:

$$\hat{f} = \arg \min_g \hat{\mathcal{R}}_t(g) \quad (7)$$

Alternatively, our goal is to rank a collection of models $\{g_i\}_{i=1,2,\dots,k}$ based on their empirical risk $\{\hat{\mathcal{R}}_t(g_i)\}_{i=1,2,\dots,k}$ in a geostatistical learning problem to aid model selection.

In order to achieve the stated goals, we need to 1) estimate the importance weights in the empirical risk approximation, and 2) remove the dependence of the approximation on a specific dataset. These two issues are addressed in the following sections.

3.1 Density Ratio Estimation

The empirical approximation of the risk $\hat{\mathcal{R}}_t(g)$, depends on estimates of the weights $\dot{w}(\mathcal{X}_u)$, which are ratios of probabilities in the target and source domains. The following problem can be posed [32]:

Definition (Density Ratio Estimation). Given two collections of samples $\{\mathcal{X}_u\}_{u \in \mathcal{D}_s}$ and $\{\mathcal{X}_v\}_{v \in \mathcal{D}_t}$ from source and target domains, estimate the density ratio $\frac{\text{Pr}_{u \in \mathcal{D}_t}(\mathcal{X})}{\text{Pr}_{u \in \mathcal{D}_s}(\mathcal{X})}$ at any new sample \mathcal{X} . In particular, estimate the ratio at all samples $\{\mathcal{X}_u\}_{u \in \mathcal{D}_s}$ from the source.

Efficient methods for density ratio estimation that perform well with high-dimensional features have been proposed in the literature. In this work we consider a fast method named Least Squares Importance Fitting (LSIF) [35, 36]. The LSIF method assumes that the weights are a linear combination of basis functions $\dot{w}(\mathcal{X}_u) = \alpha^\top \varphi(\mathcal{X}_u)$ with coefficients to be learned $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_b)$ and fixed basis $\varphi(\mathcal{X}_u) = (\varphi_1(\mathcal{X}_u), \varphi_2(\mathcal{X}_u), \dots, \varphi_b(\mathcal{X}_u))$. The LSIF estimator is derived by minimizing the squared error with the true density ratio:

$$\underset{\alpha \in \mathbb{R}^b}{\text{minimize}} \frac{1}{2} \int \left(\dot{w}(\mathcal{X}_u) - \frac{\Pr_{\mathbf{v} \in \mathcal{D}_t}(\mathcal{X}_u)}{\Pr_{\mathbf{u} \in \mathcal{D}_s}(\mathcal{X}_u)} \right)^2 \Pr_{\mathbf{u} \in \mathcal{D}_s}(\mathcal{X}_u) d\mathcal{X}_u \quad (8)$$

s.t. $\alpha \geq 0$

under the constraint that densities are always positive. By choosing b center features randomly from the target domain $\{\mathcal{X}_i\}_{i=1,2,\dots,b}$, the method introduces a Gaussian kernel basis $\varphi_i(\mathcal{X}_u) = k(\mathcal{X}_u, \mathcal{X}_i)$ that simplifies the objective function to a matrix form:

$$\underset{\alpha \in \mathbb{R}^b}{\text{minimize}} \frac{1}{2} \alpha^\top \mathbf{H} \alpha - \mathbf{h}^\top \alpha + \lambda \mathbf{1}^\top \alpha \quad (9)$$

s.t. $\alpha \geq 0$

with $\mathbf{H} = \int \varphi(\mathcal{X}_u) \varphi(\mathcal{X}_u)^\top \Pr_{\mathbf{u} \in \mathcal{D}_s}(\mathcal{X}_u) d\mathcal{X}_u$ and $\mathbf{h} = \int \varphi(\mathcal{X}_v) \Pr_{\mathbf{v} \in \mathcal{D}_t}(\mathcal{X}_v) d\mathcal{X}_v$. The regularization parameter $\lambda \geq 0$ on the coefficients α avoids overfitting, and empirical estimates of both \mathbf{H} and \mathbf{h} are easily obtained with sample averages:

$$\hat{\mathbf{H}} = \frac{1}{|\mathcal{D}_s|} \sum_{\mathbf{u} \in \mathcal{D}_s} \varphi(\mathcal{X}_u) \varphi(\mathcal{X}_u)^\top \quad (10)$$

$$\hat{\mathbf{h}} = \frac{1}{|\mathcal{D}_t|} \sum_{\mathbf{v} \in \mathcal{D}_t} \varphi(\mathcal{X}_v)$$

This quadratic optimization problem with linear inequality constraints can be solved very efficiently with modern optimization software [42, 43]. In the end, the optimal coefficients α^* are plugged back into the basis expansion for optimal estimates of the weights on new samples $\dot{w}(\mathcal{X}) = \alpha^{*\top} \varphi(\mathcal{X})$.

3.2 Weighted Cross-Validation

In order to remove the dependence of the empirical risk approximation on the dataset, we use importance-weighted cross-validation (IWCV) [30, 31]. As with the traditional cross-validation procedure, the source domain is split into k folds $\mathcal{D}_s = \cup_{j=1}^k \mathcal{D}_s^{(j)}$, and each fold $\mathcal{D}_s^{(j)}$ is hidden for error evaluation while the model $\hat{g}^{(j)}$ is learned on the remaining folds:

$$\hat{\mathcal{R}}_t^{IWCV}(\hat{g}) = \frac{1}{k} \sum_{j=1}^k \frac{1}{|\mathcal{D}_s^{(j)}|} \sum_{\mathbf{u} \in \mathcal{D}_s^{(j)}} (\dot{w}(\mathcal{X}_u))^l \hat{\mathcal{L}}(\mathcal{Y}_u, \hat{g}^{(j)}(\mathcal{X}_u)) \quad (11)$$

The main difference in the IWCV procedure are the weights that multiply each sample. The regularization exponent $l \in [0, 1]$ can be set to zero to recover the traditional estimator, or to a positive value to account for covariate shift. An optimal value for l can be found via hyperparameter search by considering another layer of cross-validation. In this work, we simply set default values for l such as $l = 1$ or $l = 0.5$.

In the rest of the paper, we combine IWCV with LSIF into a method for estimating generalization error that we term *Density*

Ratio Validation. Although IWCV is known to outperform classical cross-validation methods in non-spatial settings, little is known about its performance with geospatial data. Moreover, like all prior art, IWCV approximates the pointwise generalization error of Eq 3 as opposed to the geostatistical generalization error of Eq 2, and therefore is limited by design to non-spatial learning models.

4 EXPERIMENTS

In this section, we perform experiments to assess estimators of generalization error under varying covariate shifts and spatial correlation lengths. We consider Cross-Validation (CV), Block Cross-Validation (BCV) and Density Ratio Validation (DRV), which all rely on the same cross-validated mechanism of splitting data into folds.

First, we use synthetic Gaussian process data and simple labeling functions to construct geostatistical learning problems for which learning models have a known (via geostatistical simulation) generalization error. In this case, we assess the estimators in terms of how well they estimate the actual error under various spatial distributions. Second, we demonstrate how the estimators are used for model selection in a real application with well logs from New Zealand, which can be considered to be a dataset of moderate size in this field.

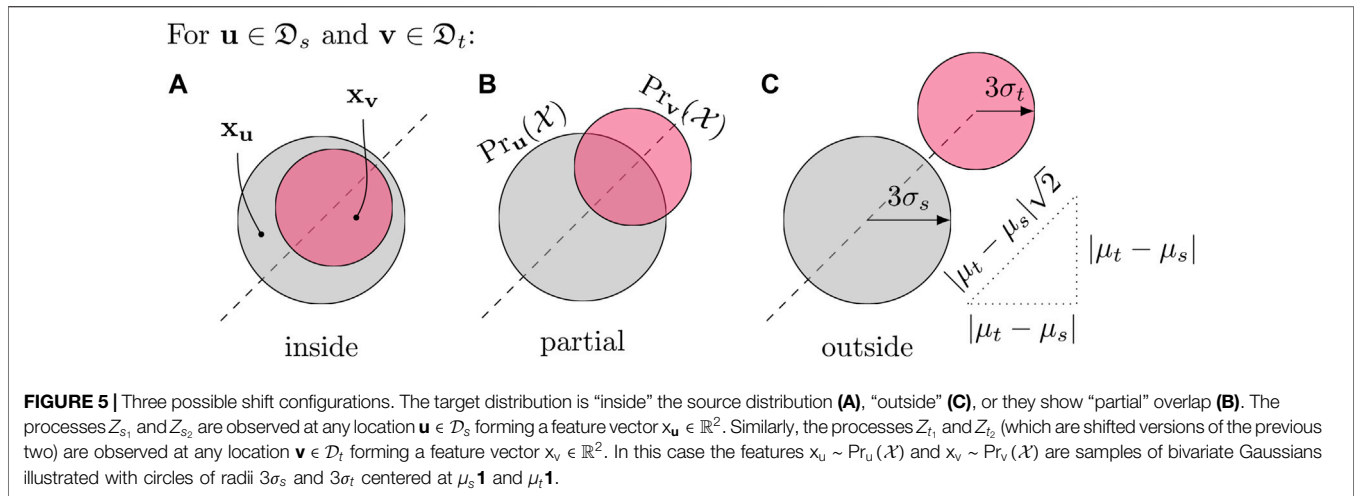
4.1 Gaussian Processes

Let Z_{s_1}, Z_{s_2} be two Gaussian processes with constant mean μ_s and variogram γ_s defined over \mathcal{D}_s , and likewise let Z_{t_1}, Z_{t_2} be two Gaussian processes with constant mean μ_t and variogram γ_t defined over \mathcal{D}_t . Denote by r_s the variogram range (or correlation length) and by σ_s^2 the variogram sill (or total variance) of the processes in the source domain. Likewise, denote by r_t and σ_t^2 the range and sill of the variogram in the target domain. It is clear that pointwise stationarity holds inside each of these domains. The feature vector $\mathcal{X}_u \in \mathbb{R}^2$ for any location $\mathbf{u} \in \mathcal{D}_s$ in the source domain has a bivariate Gaussian distribution $\mathcal{N}(\mu_s \mathbf{1}, \sigma_s^2 \mathbf{I})$, whereas the feature vector $\mathcal{X}_v \in \mathbb{R}^2$ for any location $\mathbf{v} \in \mathcal{D}_t$ in the target domain has a bivariate Gaussian distribution $\mathcal{N}(\mu_t \mathbf{1}, \sigma_t^2 \mathbf{I})$. By constraining the variogram ranges to be equal in source and target domains, that is $r_s = r_t = r$, and by requiring that both variograms pass through the origin (i.e. no nugget effect), we can investigate two types of covariate shift for various ranges r :

Mean shift. Define the shift in the mean as $\delta = c |\mu_t - \mu_s| \in [0, \infty)$ for some normalization constant $c > 0$. In this experiment, we set the value of this constant to $c = \frac{1}{3\sqrt{2}\sigma_s}$ for convenience so that a $\delta = 1$ becomes equivalent to $|\mu_t - \mu_s| = 3\sqrt{2}\sigma_s$, which in turn is equivalent to two circles (i.e. bivariate Gaussians) of radii $3\sigma_s$ touching each other along the identity line, see Figure 5.

Variance shift. Define the shift in the variance as $\tau = \sigma_t/\sigma_s \in (0, \infty)$. Here, $\tau = 1$ means absence of variance shift, $\tau < 1$ means that the variance where the model is applied is smaller than the variance where the model was trained, and $\tau > 1$ means the exact opposite of $\tau < 1$.

Geostatistical learning problems with $\tau > 1$ are very challenging to solve, and usually require additional extrapolation models, beyond the pointwise learning models discussed in this work. Therefore, we only consider cases with $\tau \leq 1$ in this experiment. More specifically, we consider all



combinations of shift in the mean and variance of Gaussian features by varying (δ, τ) in the unit square $\mathcal{B} = [0, 1] \times (0, 1]$.

Given a shift parameterized by $(\delta, \tau) \in \mathcal{B}$, we can classify it into one of three possible configurations depending on how the source and target distributions of features overlap:

$$\text{config}(\delta, \tau) = \begin{cases} \text{inside,} & 2\delta \leq 1 - \tau \\ \text{outside,} & 2\delta \geq 1 + \tau \\ \text{partial,} & \text{otherwise} \end{cases} \quad (12)$$

The first configuration in Eq. 12 refers to the case in which the target distribution $\mathcal{N}(\mu_t \mathbf{1}, \sigma_t^2 \mathbf{I})$ is “inside” the source distribution $\mathcal{N}(\mu_s \mathbf{1}, \sigma_s^2 \mathbf{I})$ meaning that the circle of radius $3\sigma_t$ centered at $\mu_t \mathbf{1}$ is contained in the circle of radius $3\sigma_s$ centered at $\mu_s \mathbf{1}$. Similarly, the second configuration refers to the case in which the target distribution is “outside” the source distribution. Finally, the third configuration refers to a “partial” overlap when the two distributions share a common set of samples but are not entirely one inside of the other. We note, however; that the illustration with circles provided in Figure 5 is only representative in the absence of spatial correlation (i.e., $r = 0$), see Figure 3.

To efficiently simulate multiple spatial samples of the processes over a regular grid domain with 100×100 locations (or pixels), we use spectral Gaussian simulation [44]. We fix the parameters of the source distribution at $\mu_s = 0$ and $\sigma_s = 1$ without loss of generality, and assume no inter-process correlation (i.e., $\rho = 0$) like we did in Figure 3. Under these modeling assumptions, we are able to investigate the spatial distribution of features as a function of shift parameters $(\delta, \tau) \in \mathcal{B}$ and variogram ranges $r \in \mathcal{C} = \{0, 10, 20\}$.

To fully specify the geostatistical learning problem, we need to specify a learning task. The task consists of predicting a binary variable \mathcal{Y}_v at locations \mathbf{v} in the target grid \mathcal{D}_t based on observations y_u of the variable at locations \mathbf{u} in the source grid \mathcal{D}_s . These observations (or labels) are synthesized using known labeling functions such as $y_u = \text{sgn}(\sin(w\|x\|_p))$, where $\|\cdot\|_p$ is the p -norm, w is the angular frequency, and sgn is the modified sign function that assigns +1 to $x \geq 0$ and -1 otherwise. The observations produced by these functions form alternating patterns in the feature space, which are not trivial to predict with simple learning models, see Figure 6. In

this experiment, we fix $p = 1$ and $w = 4$ to save computational time. Other norms and angular frequencies produce similar results.

Having defined the problem, we proceed and specify learning models in order to investigate the different estimators of generalization error. We choose two models that are based on different prediction mechanisms [2]:

Decision tree. A pointwise decision tree model \hat{f}_T makes predictions solely based on the features of the sample, without exploiting nearby features in the feature space.

K-nearest neighbors. A pointwise k-nearest neighbors model \hat{f}_N makes predictions based on nearby features, and is sometimes called a “spatial model”.

These two models \hat{f}_T and \hat{f}_N are simply representative models from the “non-spatial” and “spatial” families of models. We emphasize, however; that the term “spatial model” can be misleading in the spatial statistics literature. It is important to distinguish “spatial models” such as k-nearest neighbors that exploit the notion of proximity of features in the *feature space* from “geospatial models” that also exploit the proximity of samples in the *physical space* (or spatial domain as we have been calling it) besides their features.

The experiment proceeds as follows. For each shift $(\delta, \tau) \in \mathcal{B}$, each correlation length $r \in \mathcal{C}$, and each pointwise learning model $\hat{f} \in \{\hat{f}_T, \hat{f}_N\}$, we sample a problem $\mathcal{P}_{\delta, \tau, r}$ and estimate the generalization error of the model \hat{f} on the problem with CV, BCV and DRV. We set the hyperparameters of the CV and BCV estimators based on the fact that the correlation length never exceeds 20. For instance, we set the block side in BCV to $s = 20$, and use the equivalent number of folds in CV, i.e., $k = (100/20)^2 = 25$ for a domain with 100×100 pixels. We set the kernel width of the LSIF estimator in DRV to $\sigma = 2$ based on the synthetic Gaussian distributions, and use 10 kernels in the basis expansion. Additionally, we approximate the true generalization error of the models with Monte Carlo simulation over the target domain (e.g. 100 spatial samples).

To facilitate the visualization of the results, we introduce shift functions $\mathcal{S} : \mathcal{B} \rightarrow [0, \infty)$ that map the shift parameters (δ, τ) to a single covariate shift value, which can be interpreted loosely as the “difficulty” of the problem:

Kullback-Leibler divergence. The divergence or relative entropy between two distributions p and q is defined as

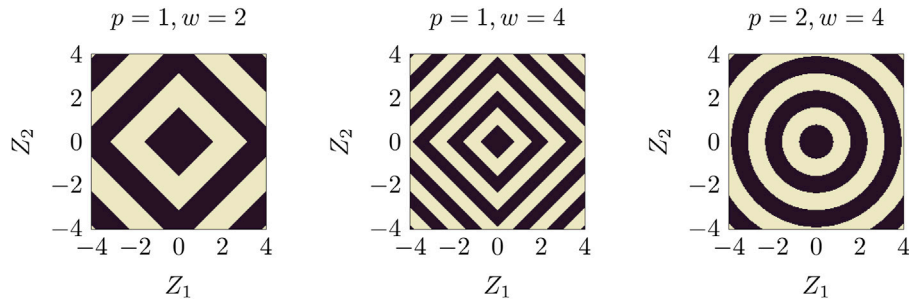


FIGURE 6 | Labeling function $f_{p,w}(x) = \text{sgn}(\sin(w|x|_p))$. Labels form alternating patterns in the feature space for different p -norm and angular frequencies w .

$S_{KL} = (p \parallel q) \int p(x) \log \frac{p(x)}{q(x)} dx$, and can be derived analytically for two (2D) Gaussian distributions $p = \mathcal{N}(\mu_t \mathbf{1}, \sigma_t^2 \mathbf{I})$ and $q = \mathcal{N}(\mu_s \mathbf{1}, \sigma_s^2 \mathbf{I})$. We derive a formula in terms of δ and τ by fixing $\sigma_s = 1$:

$$S_{KL}(\delta, \tau) = \delta^2 + \tau^2 - \log(\tau^4) - 1 \tag{13}$$

Jaccard distance. The Jaccard index between two sets A and B is defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ and the corresponding distance as $S_J(A, B) = 1 - J(A, B)$. For two (2D) Gaussian distributions, we consider A and B to be circles of radii $3\sigma_s$ and $3\sigma_t$ centered at $\mu_s \mathbf{1}$ and $\mu_t \mathbf{1}$. The distance is then expressed in terms of areas, which can be derived analytically in terms of δ and τ by fixing $\sigma_s = 1$:

$$|A| = 9\pi$$

$$|B| = 9\pi\tau^2$$

$$C_1 = 9 \arccos\left(\frac{2\delta^2 + 9(1 - \tau^2)}{6\sqrt{2}\delta}\right)$$

$$C_2 = 9 \tau^2 \arccos\left(\frac{2\delta^2 + 9(\tau^2 - 1)}{6\sqrt{2}\delta\tau}\right) \tag{14}$$

$$C_3 = \frac{1}{2} \sqrt{(9(1 + \tau)^2 - 2\delta^2)(2\delta^2 - 9(1 - \tau)^2)}$$

$$|A \cap B| = C_1 + C_2 - C_3$$

$$|A \cup B| = |A| + |B| - |A \cap B|$$

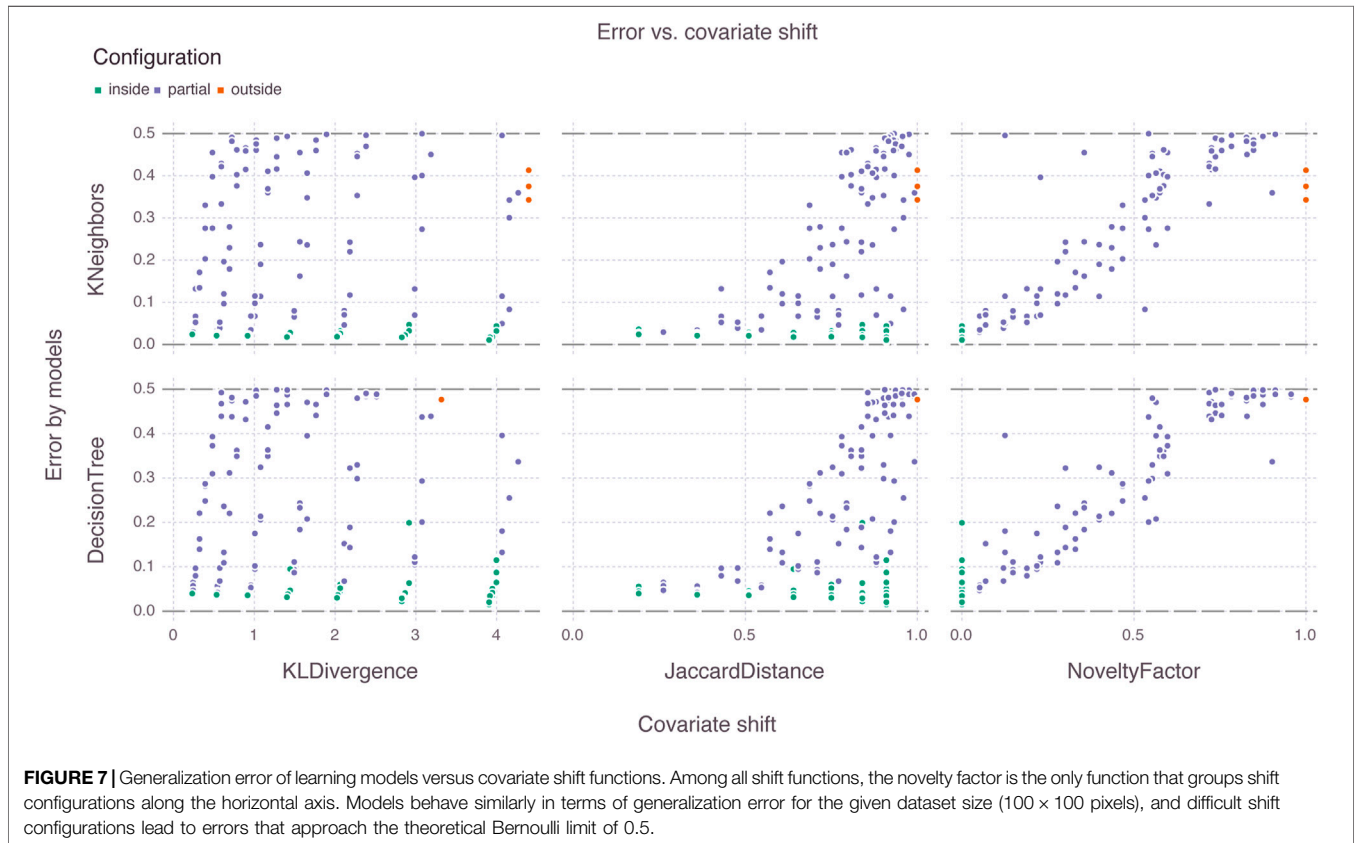


FIGURE 7 | Generalization error of learning models versus covariate shift functions. Among all shift functions, the novelty factor is the only function that groups shift configurations along the horizontal axis. Models behave similarly in terms of generalization error for the given dataset size (100 × 100 pixels), and difficult shift configurations lead to errors that approach the theoretical Bernoulli limit of 0.5.

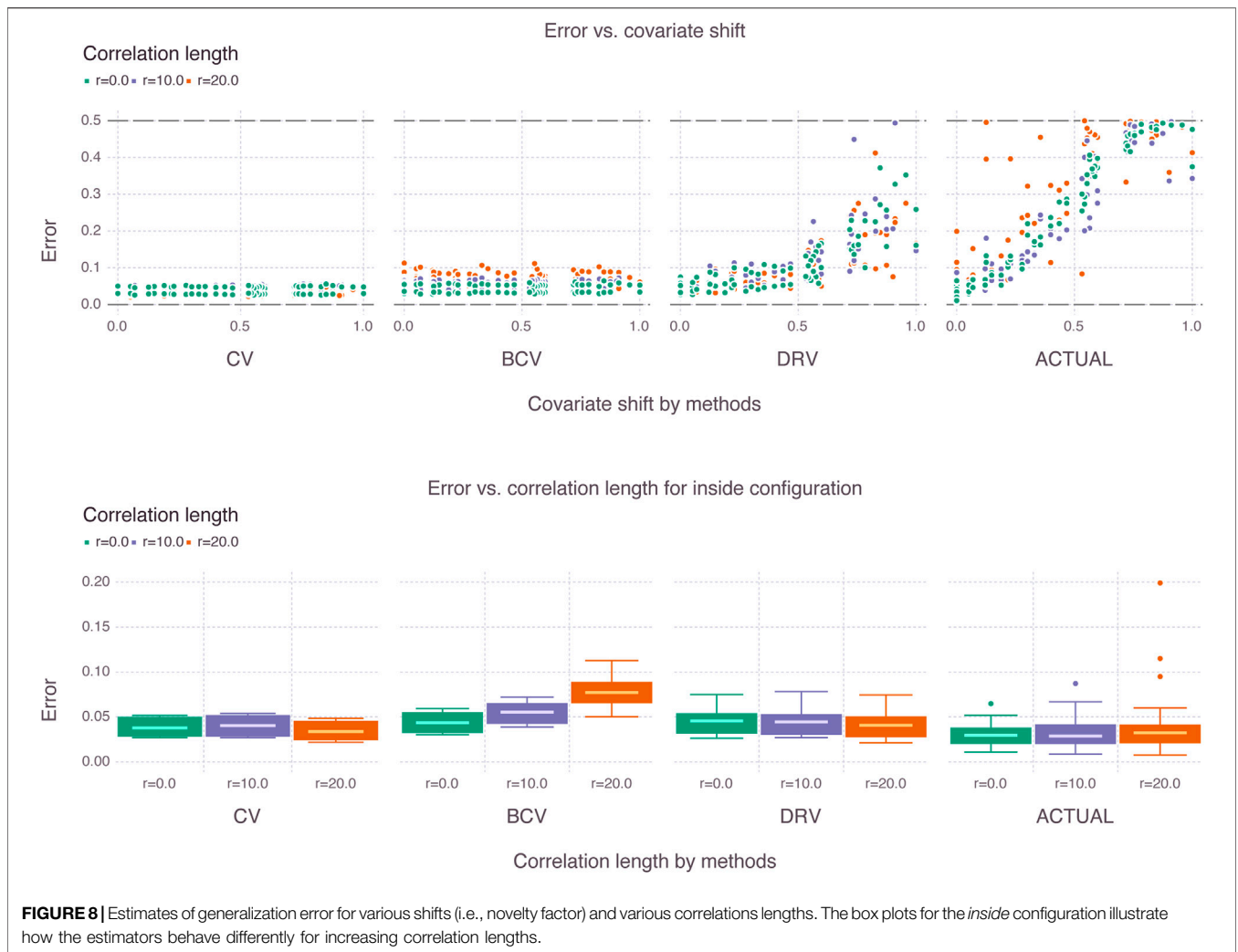


FIGURE 8 | Estimates of generalization error for various shifts (i.e., novelty factor) and various correlations lengths. The box plots for the *inside* configuration illustrate how the estimators behave differently for increasing correlation lengths.

Novelty factor. We propose a new shift function termed the *novelty factor* inspired by the geometric view of Jaccard. First, we define the novelty of B with respect to A as $N(B/A) = \frac{|B-A \cap B| - |A \cap B|}{|B|}$, and notice that it is the fraction of B that is outside of A minus the fraction of B that is inside of A . Second, we restrict the definition to cases with $|B| \leq |A|$ (e.g. Gaussian case with $\tau \leq 1$), and notice that the novelty $N(B/A)$ lies in the interval $[-1, 1]$. Finally, we define the novelty factor $S_N(A, B) = \frac{N(B/A)+1}{2}$ in the interval $[0, 1]$, which can be easily computed for the Gaussian case using the formulas derived in Eq 14.

We plot the true generalization error of the models as a function of the different covariate shifts in Figure 7, and color the points according to their shift configuration (see Eq 12). In this plot, the horizontal dashed line intercepting the vertical axis at 0.5 represents a model that assigns positive and negative labels to samples at random with equal weight (i.e., Bernoulli variable).

Among the three shift functions, the novelty factor is the only function that groups shift configurations along the horizontal axis. In this case, configurations deemed easy (i.e., where the target distribution is *inside* the source distribution) appear first, then configurations of moderate difficulty (i.e., *partial* overlap) appear next, and finally difficult configurations (i.e., target is *outside* the source) appear near

the theoretical Bernoulli limit. The Kullback-Leibler divergence and the Jaccard distance fail to summarize the shift parameters into a one-dimensional visualization, and are therefore omitted in the next plots.

The two models behave similarly in terms of generalization error for the given dataset size (i.e., 100×100 pixels), and therefore can be aggregated into a single scatter to increase the confidence in the observed trends.

We plot the CV, BCV and DRV estimates of generalization error versus covariate shift (i.e., novelty factor) in the top row of Figure 8, and color the points according to their correlation length. We omit a few DRV estimates that suffered from numerical instability in challenging *partial* or *outside* configurations, and show the box plot of error estimates in the bottom row of the figure.

First, we emphasize that the CV and BCV estimates remain constant as a function of covariate shift. This is expected given that these estimators do not make use of the target distribution. The DRV estimates increase with covariate shift as expected, but do not follow the same rate of increase of the true (or actual) generalization error obtained with Monte Carlo simulation. Second, we emphasize in the box plots for the *inside* configuration that the correlation length affects the estimators

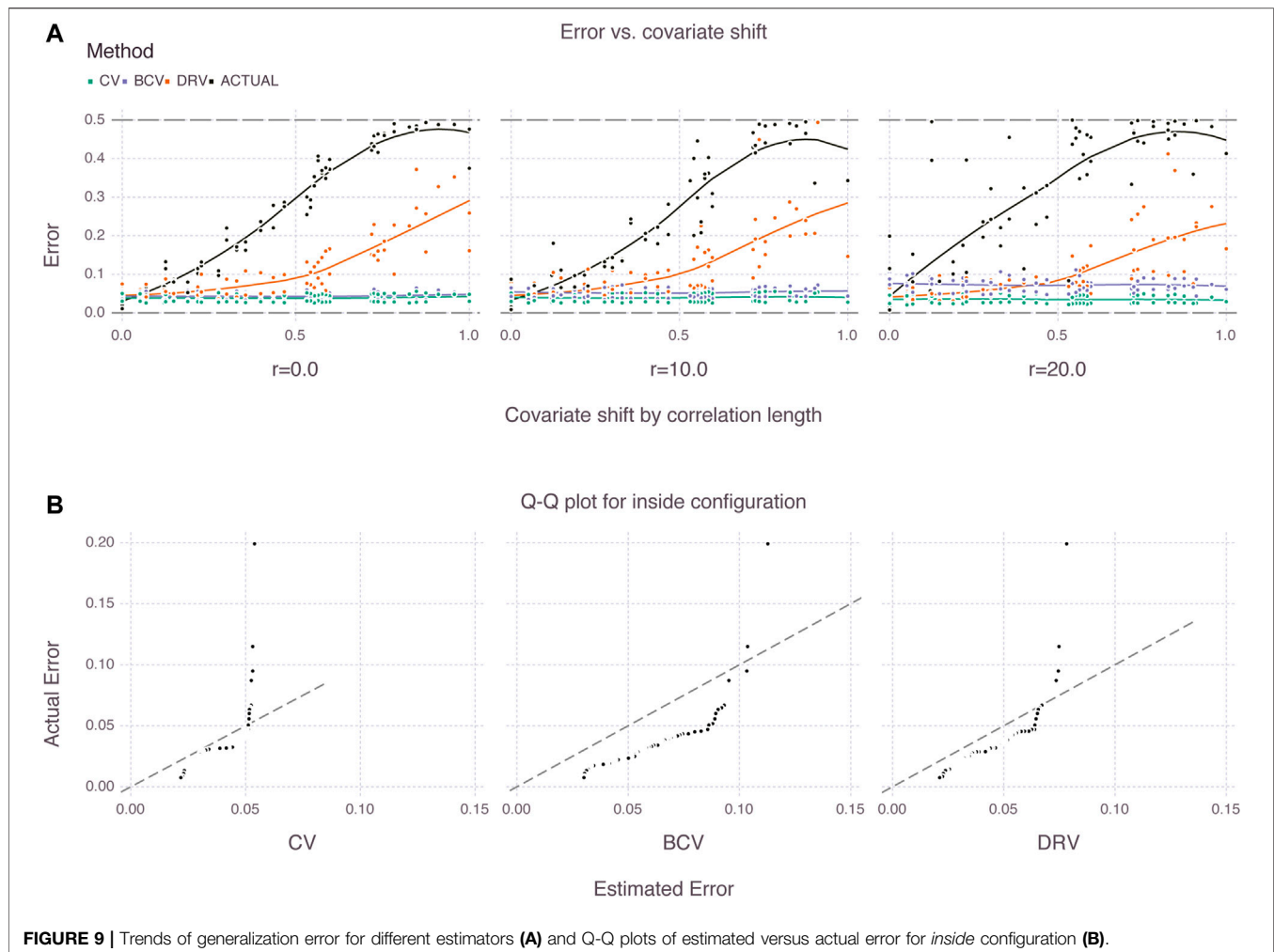


FIGURE 9 | Trends of generalization error for different estimators **(A)** and Q-Q plots of estimated versus actual error for *inside* configuration **(B)**.

differently. The CV estimator becomes more optimistic with increasing correlation length, whereas the BCV estimator becomes less optimistic, a result that is also expected from prior art. Additionally, the interquartile range of the BCV estimator increases with correlation length. It is not clear from the box plots that a trend exists for the DRV estimator. The actual generalization error behaves erratically in the presence of large correlation lengths as indicated by the scatter and box plots.

In order to better visualize the trends in the estimates, we smooth the scatter plots with locally weighted regression per correlation length in the top row of **Figure 9**, and show in the bottom row of the figure the Q-Q plots of the different estimates against the actual generalization error for the *inside* configuration where all estimators are supposed to perform well.

From the figure, there exists a gap between the DRV estimates and the actual generalization error of the models for all covariate shifts. This gap is expected given that the target distribution may be very different from the source distribution, particularly in *partial* or *outside* shift configurations. On the other hand, the gap also seems to be affected by the correlation length, and is largest with 20 pixels of correlation. Additionally, we emphasize in the Q-Q plots that the BCV estimates are biased due to the systematic selection of folds. The BCV estimates

are less optimistic than the CV estimates, which is a desired property in practice, however there is no guarantee that the former estimates will approximate well the actual generalization error of the models.

4.2 New Zealand Dataset

Unlike the previous experiment with synthetic Gaussian process data and known generalization error, this experiment consists of applying the CV, BCV and DRV estimators to a real dataset of well logs prepared in-house [37]. We quickly describe the dataset, introduce the related geostatistical learning problems, and use error estimates to rank learning models. Finally, we compare these ranks with an ideal rank obtained with additional label information that is not available during the learning process.

The dataset consists of 407 wells in the Taranaki basin, including the main geophysical logs and reported geological formations. The basin comprises an area of about 330.000 km², located broadly onshore and offshore the New Zealand west coast (see **Figure 10**). Well trajectories are georeferenced in UTM coordinates (X and Y) and true vertical depth (Z).

We split the wells into onshore and offshore locations in order to introduce a geostatistical learning problem with covariate shift.

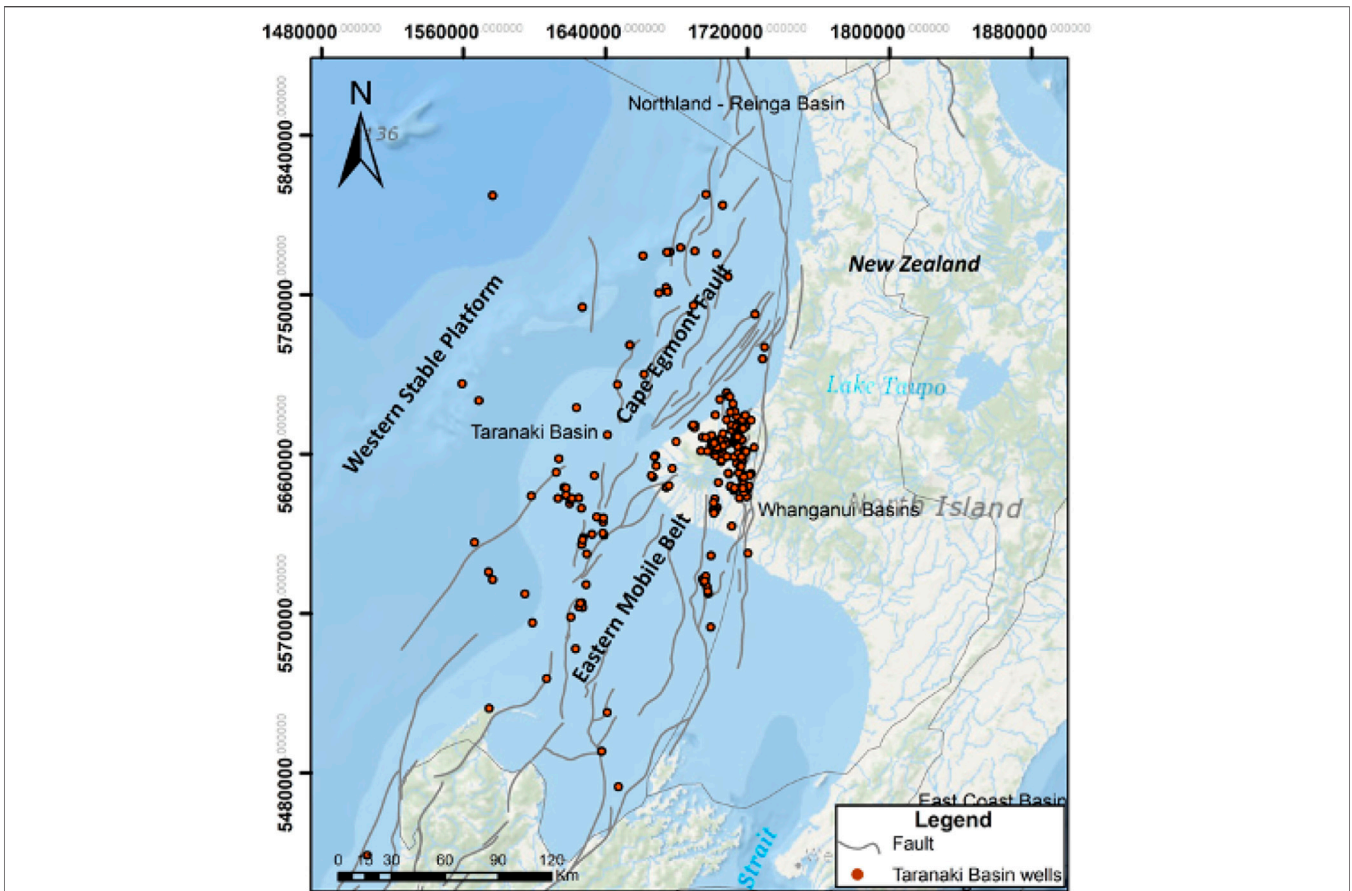


FIGURE 10 | Curated dataset with 407 wells in the Taranaki basin, New Zealand. The basin comprises an area of about 330.000 km², located broadly onshore and offshore the New Zealand west coast.

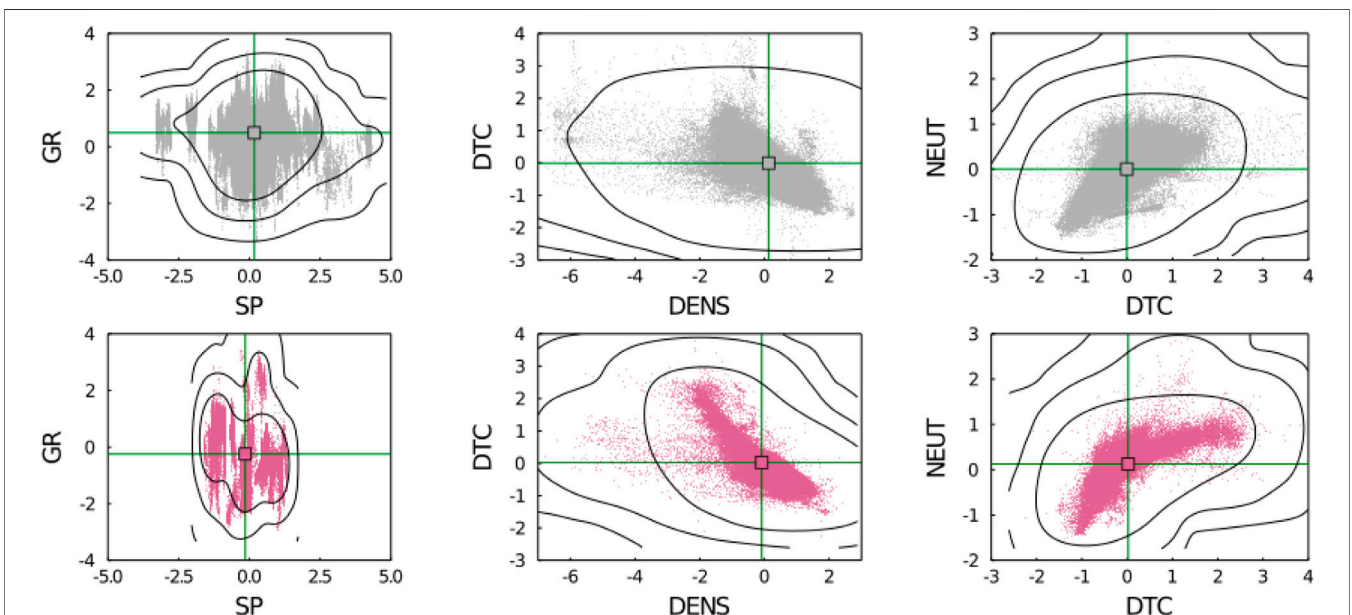


FIGURE 11 | Distribution of main geophysical logs onshore (gray) and offshore (purple) centered by the mean and divided by the standard deviation. Visible covariate shift in the scatter and contour plots.

TABLE 1 | Estimates of generalization error with different estimators for the onshore-to-offshore problem. The CV estimator produces estimates that are the most distant to the actual target error due to covariate shift and spatial correlation. None of the estimators is capable of ranking the models correctly. They all select complex models with low generalization ability.

MODEL	SOURCE	TARGET	CV	BCV	DRV
Ridge	0.262	0.291	0.263	0.326	0.278
Logistic	0.262	0.294	0.262	0.322	0.275
KNeighbors	0.023	0.355	0.034	0.227	0.043
GaussianNB	0.256	0.321	0.256	0.288	0.284
BayesianLDA	0.262	0.291	0.261	0.326	0.294
Perceptron	0.375	0.395	0.344	0.369	0.392
DecisionTree	0.0	0.378	0.034	0.249	0.043
Dummy	0.499	0.497	0.499	0.502	0.524
TARGET RANK	CV RANK	BCV RANK	DRV RANK		
BayesianLDA	KNeighbors	KNeighbors	KNeighbors		
Ridge	DecisionTree	DecisionTree	DecisionTree		
Logistic	GaussianNB	GaussianNB	Logistic		
GaussianNB	BayesianLDA	Logistic	Ridge		
KNeighbors	Logistic	Ridge	GaussianNB		
DecisionTree	Ridge	BayesianLDA	BayesianLDA		
Perceptron	Perceptron	Perceptron	Perceptron		
Dummy	Dummy	Dummy	Dummy		

TABLE 2 | Estimates of generalization error with different estimators for the problem without covariate shift. The BCV estimator produces estimates that are the most distant to the actual target error due to bias from its systematic selection of folds. All estimators are capable of ranking the models in the absence of covariate shift.

MODEL	SOURCE	TARGET	CV	BCV	DRV
Ridge	0.267	0.273	0.269	0.329	0.259
Logistic	0.266	0.271	0.265	0.324	0.254
KNeighbors	0.026	0.04	0.04	0.232	0.042
GaussianNB	0.265	0.268	0.265	0.308	0.262
BayesianLDA	0.267	0.273	0.266	0.329	0.264
Perceptron	0.445	0.442	0.362	0.402	0.392
DecisionTree	0.0	0.039	0.039	0.244	0.041
Dummy	0.499	0.498	0.499	0.497	0.476
TARGET RANK	CV RANK	BCV RANK	DRV RANK		
DecisionTree	DecisionTree	KNeighbors	DecisionTree		
KNeighbors	KNeighbors	DecisionTree	KNeighbors		
GaussianNB	GaussianNB	GaussianNB	Logistic		
Logistic	Logistic	Logistic	Ridge		
Ridge	BayesianLDA	BayesianLDA	GaussianNB		
BayesianLDA	Ridge	Ridge	BayesianLDA		
Perceptron	Perceptron	Perceptron	Perceptron		
Dummy	Dummy	Dummy	Dummy		

The problem consists of predicting the rock formation from well logs offshore after learning a model with well logs and reported (i.e., manually labeled) formations onshore. The well logs considered are gamma ray (GR), spontaneous potential (SP), density (DENS), compressional sonic (DTC) and neutron porosity (NEUT). We eliminate locations with missing values for these logs and investigate a balanced dataset with the two most frequent formations—Urenui and Manganui. We normalize the logs and illustrate the covariate shift property by comparing the scatter plots of onshore and offshore locations in **Figure 11**. Additionally, we define a second geostatistical learning problem without covariate shift. In this case, we join all locations filtered in the previous problem and sample two new sets of locations with

sizes respecting the same source-to-target proportion (e.g., 300000 : 50000).

We set the hyperparameters of the error estimators based on variography and according to available computational resources. In particular, we set blocks for the BCV estimators with sides $10000 \times 10000 \times 500$ that are much greater than the vertical and horizontal correlation lengths estimated from empirical variograms. We obtain the corresponding number of folds $k = 99$ for the CV estimator by partitioning the bounding box of onshore wells into blocks with the given sides. Similarly to the previous experiment with synthetic Gaussian process data, we set the kernel width in DRV to $\sigma = 2$ given that the well logs were normalized to have unit variance. Finally, we select a list of learning models to rank including Ridge classification

(Ridge), logistic regression (Logistic), k-nearest neighbors (KNeighbors), naive Bayes (GaussianNB), linear discriminant analysis (BayesianLDA), perceptron (Perceptron), decision tree (DecisionTree), and a dummy model that reproduces the marginal distribution of formations in the source domain (Dummy).

In **Table 1**, we report the results for the onshore-to-offshore problem. In the upper part of the table we compare side-by-side the error estimates of the different methods. We highlight the closest estimates to the target error in blue color, and the most distant in red color. We emphasize that the target error is the error of the model in one single realization of the process, and is *not* the generalization error averaged over multiple spatial realizations. In spite of this important distinction, we still think it is valuable to compare it with the estimates of generalization error given by CV, BCV and DRV since these methods were all derived under pointwise learning assumptions, and are therefore smooth averages over multiple points exactly like the error estimated from the single realization of the target. In the bottom part of the table, we report the model ranks derived from the error estimates as well as the ideal rank derived from the target error.

Among the three estimators of generalization error, the CV estimator produces estimates that are the most distant from the target error, with a tendency to underestimate the error. The BCV estimator produces estimates that are higher than the CV estimates, and consequently closer to the target error in this case. The DRV estimator produces the closest estimates for most models, however; like the CV estimator it fails to approximate the error for models like KNeighbors and DecisionTree that are over-fitted to the source distribution. The three estimators fail to rank the models under covariate shift and spatial correlation. Over-fitted models with low generalization ability are incorrectly ranked at the top of the list, and the best models, which are simple “linear” models, appear at the bottom. We compare these results with the results obtained for the problem without covariate shift in **Table 2**.

From **Table 2**, the CV estimator produces estimates that are the closest to the target error. The BCV estimator produces estimates that are higher than the CV estimates as before, however this time this means that the BCV estimates are the most distant to the target error. The DRV estimator produces estimates that are not the closest nor the most distant to the target error. The three estimators successfully rank the models from simple linear models at the bottom of the list to more complex learning models at the top. Unlike the previous problem with covariate shift, this time complex models like KNeighbors and DecisionTree show high generalization ability.

5 CONCLUSION

In this work, we introduce *geostatistical (transfer) learning*, and demonstrate how most prior art in statistical learning with geospatial data fits into a category we term *pointwise learning*. We define geostatistical generalization error and demonstrate how existing estimators from the spatial statistics literature such as block cross-validation are derived for that specific category of learning, and are therefore unable to account for general spatial errors.

We propose experiments with spatial data to compare estimators of generalization error, and illustrate how these

estimators fail to rank models under covariate shift and spatial correlation. Based on the results of these experiments, we share a few remarks related to the choice of estimators in practice:

- The apparent quality of the BCV estimator is falsified in the Q-Q plots of **Figure 9** and in **Table 2**. The systematic bias produced by the blocking mechanism only guarantees that the error estimates are higher than the CV estimates. When the CV estimates are good (i.e. no covariate shift), the BCV estimates are unnecessarily pessimistic.
- The CV estimator is not adequate for geostatistical learning problems that show various forms of covariate shift. Situations without covariate shift are rare in geoscientific settings, and since the DRV estimator works reasonably well for both situations (i.e. with and without shift), it is recommended instead.
- Nevertheless, both the CV and DRV estimators suffer from a serious issue with over-fitted models in which case they largely underestimate the generalization error. For risk-averse applications where one needs to be careful about the generalization error of the model, the BCV estimator can provide more conservative results.
- None of the three estimators were capable of ranking models correctly under covariate shift and spatial correlation. This is an indication that one needs to be skeptical about interpreting similar rankings available in the literature.

Finally, we believe that this work can motivate methodological advances in learning from geospatial data, including research on new estimators of geostatistical generalization error as opposed to pointwise generalization error, and more explicit treatments of spatial coordinates of samples in learning models.

Computer Code Availability

All concepts and methods developed in this paper are made available in the GeoStats.jl project [45]. The project is hosted on GitHub under the MIT¹ open-source license: <https://github.com/JuliaEarth/GeoStats.jl>.

Experiments of this specific work can be reproduced with the following scripts: <https://github.com/IBM/geostats-gen-error>.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in the same repository with the code: <https://github.com/IBM/geostats-gen-error>.

AUTHOR CONTRIBUTIONS

JH: Conceptualization, methodology, software, formal analysis, investigation, visualization, writing—original draft; MZ: Methodology, validation; BC: Data curation, validation; BZ: Methodology, validation, supervision.

¹<https://opensource.org/licenses/MIT>

REFERENCES

1. Vapnik V. *Statistical Learning Theory*. New York: Wiley (1998).
2. Hastie T, Tibshirani R, and Friedman J. *Elements of Statistical Learning*. 2nd ed. (2009). doi:10.1007/978-0-387-84858-7
3. Mitchell TM. *Machine Learning*. New York: McGraw-Hill (1997).
4. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Machine Learn Res* (2011) 12:2825–2830.
5. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wickes M, Yu Y, and Zheng X. TensorFlow: A System for Large-Scale Machine Learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. OSDI (2016). arXiv:1605.08695.
6. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In: H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, and R Garnett, editors. *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc. (2019). p. 8024–35.
7. Innes M. Flux: Elegant Machine Learning with Julia. *Joss* (2018) 3:602. doi:10.21105/joss.00602
8. Arlot S, and Celisse A. A Survey of Cross-Validation Procedures for Model Selection. *Statist Surv*. (2010). 4: 4728, 2010 . arXiv:0907. doi:10.1214/09-SS054
9. Ferracioli MA, Bocca FF, and Rodrigues LHA. Neglecting Spatial Autocorrelation Causes Underestimation of the Error of Sugarcane Yield Models. *Comput Electron Agric* (2019) 161:233–40. doi:10.1016/j.compag.2018.09.003
10. Jiang Y, Krishnan D, Mobahi H, and Bengio S. Predicting the Generalization Gap in Deep Networks with Margin Distributions. 7th International Conference on Learning Representations. ICLR (2019).arXiv:1810.00113.
11. Zhang C, Recht B, Bengio S, Hardt M, and Vinyals O. Understanding Deep Learning Requires Rethinking Generalization. In: 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings (2019).
12. Vehtari A, and Ojanen J. A Survey of Bayesian Predictive Methods for Model Assessment, Selection and Comparison. *Statist Surv* 6. (2012) 03530, 2012 . arXiv:1611. doi:10.1214/12-ss102
13. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J R Stat Soc Ser B (Methodological)* (1974) 36(2):111–33. doi:10.1111/j.2517-6161.1974.tb00994.x
14. Geisser S. The Predictive Sample Reuse Method with Applications. *J Am Stat Assoc* (1975) 70:320–8. doi:10.1080/01621459.1975.10479865
15. Burman P. A Comparative Study of Ordinary Cross-Validation, V-fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika* (1989) 76: 503–14. doi:10.1093/biomet/76.3.503
16. Cracknell MJ, and Reading AM. Geological Mapping Using Remote Sensing Data: A Comparison of Five Machine Learning Algorithms, Their Response to Variations in the Spatial Distribution of Training Data and the Use of Explicit Spatial Information. *Comput Geosciences* (2014) 63:22–33. doi:10.1016/j.cageo.2013.10.008
17. Baglaeva EM, Sergeev AP, Shichkin AV, and Buevich AG. The Effect of Splitting of Raw Data into Training and Test Subsets on the Accuracy of Predicting Spatial Distribution by a Multilayer Perceptron. *Math Geosci* (2020) 52:111–21. doi:10.1007/s11004-019-09813-9
18. Burman P, Chow E, and Nolan D. A Cross-Validatory Method for Dependent Data. *Biometrika* (1994) 81:351–8. doi:10.1093/biomet/81.2.351
19. Le Rest K, Pinaud D, Monestiez P, Chadoeuf J, and Bretagnolle V. Spatial Leave-One-Out Cross-Validation for Variable Selection in the Presence of Spatial Autocorrelation. *Glob Ecol Biogeogr* (2014) 23:811–20. doi:10.1111/geb.12161
20. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillaer-Arroita G, et al. Cross-validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure *Ecography* (2017) 40:913–29. doi:10.1111/ecog.02881
21. Pohjankukka J, Pahikkala T, Nevalainen P, and Heikkonen J. Estimating the Prediction Performance of Spatial Models via Spatial K-fold Cross Validation. *Int J Geographical Inf Sci* (2017) 31:2001–19. doi:10.1080/13658816.2017.1346255
22. Chilès J-P, and Delfiner P. *Geostatistics: Modeling Spatial Uncertainty*. Hoboken, NJ: Wiley (2012).
23. Pan SJ, and Yang Q. A Survey on Transfer Learning. *IEEE Trans Knowl Data Eng* (2010) 22:1345–59. doi:10.1109/TKDE.2009.191
24. Weiss K, Khoshgoftaar TM, and Wang D. A Survey of Transfer Learning. *J Big Data* (2016) 3. doi:10.1186/s40537-016-0043-6
25. Silver DL, and Bennett KP. Guest Editor's Introduction: Special Issue on Inductive Transfer Learning. *Mach Learn* (2008) 73:215–20. doi:10.1007/s10994-008-5087-1
26. Zadrozny B, Langford J, and Abe N. Cost-sensitive Learning by Cost-Proportionate Example Weighting. In: Proceedings - IEEE International Conference on Data Mining . ICDM (2003). doi:10.1109/icdm.2003.1250950
27. Zadrozny B. Learning and Evaluating Classifiers under Sample Selection Bias. In: Proceedings, Twenty-First International Conference on Machine Learning. ICML (2004). doi:10.1145/1015330.1015425
28. Wei Fan W, Davidson I, Zadrozny B, and Philip S. Yu PS. An Improved Categorization of Classifiers Sensitivity on Sample Selection Bias. In: Proceedings - IEEE International Conference on Data Mining. ICDM (2005). doi:10.1109/ICDM.2005.24
29. Joaquin Q-C, Sugiyama M, Schwaighofer A, and Lawrence ND. *Dataset Shift in Machine Learning*. Cambridge, MA: The MIT Press (2008).
30. Sugiyama M, Blankertz B, Krauledat M, Dornhege G, and Müller K-R. Importance-weighted Cross-Validation for Covariate Shift. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)(2006). p. 354–63. doi:10.1007/11861898_36
31. Sugiyama M, Krauledat M, and Müller KR. Covariate Shift Adaptation by Importance Weighted Cross Validation. *J Machine Learn Res* (2007) 8:985–1005.
32. Sugiyama M, Suzuki T, and Kanamori T. Density Ratio Estimation in Machine Learning (2012). doi:10.1017/CBO9781139035613
33. Huang J, Smola AJ, Gretton A, Borgwardt KM, and Schölkopf B. Correcting Sample Selection Bias by Unlabeled Data. In Advances in Neural Information Processing Systems (2007). doi:10.7551/mitpress/7503.003.0080
34. Sugiyama M, Nakajima S, Kashima H, Von Bünau P, and Kawanabe M. Direct Importance Estimation with Model Selection and its Application to Covariate Shift Adaptation. In Advances in Neural Information Processing Systems (2009).
35. Kanamori T, Hido S, and Sugiyama M. Efficient Direct Density Ratio Estimation for Non-stationarity Adaptation and Outlier Detection. In Advances in Neural Information Processing Systems (2009).
36. Kanamori T, Hido S, and Sugiyama M. A Least-Squares Approach to Direct Importance Estimation. *J Machine Learn Res* (2009) 10:1391–1445.
37. Carvalho BW, Oliveira M, Avalone M, Hoffmann J, Szwarcman D, Guevara Diaz J, et al. Taranaki basin Curated Well Logs. *Zenodo* (2020). doi:10.5281/zenodo.383295510.5281/zenodo.3832955
38. Hoffmann J, and Zadrozny B. Efficient Variography with Partition Variograms. *Comput Geosciences* (2019) 131:52–9. doi:10.1016/j.cageo.2019.06.013
39. Alabert F. The Practice of Fast Conditional Simulations through the LU Decomposition of the Covariance Matrix. *Math Geol* (1987) 19:369–86. doi:10.1007/BF00897191
40. Hoffmann J, Scheidt C, Barfod A, and Caers J. Stochastic Simulation by Image Quilting of Process-Based Geological Models. *Comput Geosciences* (2017) 106: 18–32. doi:10.1016/j.cageo.2017.05.012
41. Mariethoz G, Renard P, and Straubhaar J. The Direct Sampling Method to Perform Multiple-point Geostatistical Simulations. *Water Resour Res* (2010) 46. doi:10.1029/2008WR007621
42. K Mogensen P, and N Riseth A. Optim: A Mathematical Optimization Package for Julia. *Joss* (2018) 3:615. doi:10.21105/joss.00615
43. Dunning J, Huchette J, and Lubin M. JuMP: A Modeling Language for Mathematical Optimization. *SIAM Rev* (2017) 59:295–320. arXiv:1508. doi:10.1137/15M1020575
44. Gutjahr A, Bullard B, and Hatch S. General Joint Conditional Simulations Using a Fast Fourier Transform Method. *Math Geol* (1997) 29:361–89. doi:10.1007/BF02769641
45. Hoffmann J. GeoStats.jl - High-Performance Geostatistics in Julia. *Joss* (2018) 3:692. doi:10.21105/joss.0069210.21105/joss.00692

Conflict of Interest: All authors were employed by the company IBM Research.

Copyright © 2021 Hoffmann, Zortea, de Carvalho and Zadrozny. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.