This is the final peer-reviewed accepted manuscript of:

Vagheggini, A., Bruno, F., and Cocchi, D. (2016) A competitive design-based spatial predictor. Environmetrics, 27: 454– 465. doi: 10.1002/env.2423

The final published version is available online at: https://doi.org/10.1002/env.2423

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Environmetrics 00, 1–24

DOI: 10.1002/env.XXXX

A competitive design-based spatial predictor

A. Vaggeggini[a]* F. Bruno[a] and D. Cocchi[a]

Summary: Under the finite population design-based framework, locations' spatial information coordinates of a population have traditionally been used to develop efficient sampling designs rather than for estimation or prediction. We propose to enhance design-based individual prediction by exploiting the spatial information derived from geography, which is available for each population element before sampling. Individual predictors are obtained by reinterpreting deterministic interpolators under the finite population design-based framework, making it possible to derive their statistical properties. Monte Carlo experiments on real and simulated data help to appreciate the performances of the proposed approach in comparison both with estimators that do not employ spatial information and with kriging. We found that under the most favourable conditions for kriging, the proposed predictor shows good performances, particularly for small sample sizes.

Keywords: Finite populations design-based inference; Finite populations model-based inference; Individual spatial prediction.

1. INTRODUCTION

Prediction in finite populations is often performed under a model-based approach, where a superpopulation is assumed (Little, 2004, 2014). Samples are used for estimating model hyperparameters, which are the basis for the prediction of unsampled individual values (Bolfarine and Zacks, 1992). Model misspecification is, in this case, a real danger: the unsuitability of the proposed superpopulation model is a major criticism to model-based inference in finite populations. Superpopulation models are also the typical statistical

[a] *Department of Statistical Sciences, University of Bologna*

[*] *Correspondence to: Department of Statistical Sciences, University of Bologna, via delle Belle Arti 41 40126, Bologna, Italy. E-mail: alessandr.vagheggin2@unibo.it*

tools for individual probabilistic prediction in spatially structured populations; the most popular method is kriging, which exploits a stochastic setting opposite to the deterministic interpolators that are popular in geography.

On the other hand design-based inference has a very long and established tradition in finite population theory. It relies on considering the population elements as unknown non-random quantities, while the only source of randomness lies on the discrete distribution that manages sample drawing (Gregoire, 1998). Totals, means or ratios are the most common objects of estimation. Design-based methods have traditionally not been considered suitable for spatially structured populations due to a conceptual misunderstanding first highlighted by Brus and de Gruijter (1997). Indeed, probabilistic sampling was believed not able to capture the spatial dependence of the problem at hand. However, in recent decades, a reappraisal of these methods has been promoted (Cox *et al.*, 1997; Stehman, 2000; Wang *et al.*, 2012) because, as suggested by de Gruijter and ter Braak (1990), a probabilistic sampling design selects locations and not the values of the variable under study. Baffetta *et al.* (2009) derive a design-based estimator of population synthetic quantities based on the k-nearest neighbour method for data from remotely sensed digital imagery. Cicchitelli and Montanari (2012) use geographical relationships as auxiliary variables in the structure of design-based estimators, allowing them to sensibly improve the performance of the estimators of the population total. Margalho *et al.* (2014) instead propose to include the sampling design in the prediction model as a covariate.

Individual spatial prediction is routinely performed either under a statistical model-based framework, usually via kriging (Cressie, 1993; Stein, 1999), or via deterministic interpolators in a non-stochastic framework (Mitas and Mitasova, 1999). Both approaches present some drawbacks. Indeed, kriging may not be suitable when the available sample size is small because it needs the semivariogram parameter estimation, performed using the few locations available and therefore only a few spatial lags. Conversely, deterministic spatial interpolators which use weights that depend on functions of the spatial distances between the observed locations, fail to capture the random nature of the phenomenon being studied, because they are based only on the geographical arrangement, making it impossible to assess any probabilistic property (Webster and Oliver, 2007; Li and Heap, 2008).

In this paper, we illustrate how probabilistic structures can be associated with such interpolators. Thus, when the set of available locations is seen as the outcome of a probabilistic sampling design, any deterministic interpolator can be reinterpreted under randomization. This consideration fits with the finite population design-based inference paradigm, where the weighting system associated with any estimator is constructed for the whole population before sampling and is not constructed only for the available data set. Recently, Vagheggini (2013) proposed a new method for design-based spatial inference, leading to a predictor able to use the spatial information known at the population level to obtain individual values at any location.

Our aim is to use spatial information at the estimation level rather than at the sampling design level. When dealing with spatial data under a design-based framework, we consider geographical locations as further information associated with each population element. Spatial coordinates constitute a very special type of information ; geographical proximity can improve inference, but its influence must be assessed. Spatial coordinates must be kept separate from labels, still remaining population elements identifiers; thus, when coordinates are preserved, simple random sampling, which is the basic randomization of a population, does not contrast with spatial inference. This setting allows for a renewed motivation to equal probability random sampling, in a context where spatial information is separated from labels. On the converse, varying probability sampling is able to exploit information on an auxiliary variable that customarily does not correspond to spatial information. Its degree of association with study variable is often measured by the linear correlation coefficient.

In what follows, we refine the approach of Bruno *et al.* (2013) under a probabilistic sampling design. The presentation is structured as follows. In the next section, spatial interpolators are revisited under the design-based framework. Section 3 summarizes the statistical properties of the individual predictor. Monte Carlo experiments on a real environmental data and simulated populations are described in Section 4, where we compare kriging with the results obtained from the proposed design-based strategy. A thorough discussion concludes the paper. Appendix A contains proofs of the propositions stated in Section 3 and Appendix B reports results in the special case of simple random sampling.

2. REVISITING SPATIAL INTERPOLATION UNDER THE DESIGN-BASED FRAMEWORK

In this work, we introduce a methodology for design-based individual prediction. Starting from a deterministic interpolator, we consider the set of known locations used to construct the interpolating function as the realization of a fixed-size probabilistic sampling design. We thus need to conform the usual design-based inference assumptions to the case under study. First, design-based techniques rely upon considering the values in the population as fixed but unknown; therefore, the probabilistic sampling design is the only source of randomness. For this reason, without loss of generality, we assume that the values at any location in the domain are the outcome of an unknown deterministic function of the coordinates. Second, in this work, we specifically address the case of a finite population; however, following the results of Cordy (1993), extending this case to the case of sampling from a continuous domain is straightforward (Vagheggini, 2013).

We develop the idea of using the spatial information at the estimation level rather than at the sampling design level, as already introduced by Bruno *et al.* (2013). The starting point is a deterministic spatial interpolator that usually is not associated with any uncertainty measure. When the set of known locations is considered as a realization of a probabilistic sampling design, then the deterministic interpolator itself can be randomized because the extraction of each population element is managed by probability.

2.1. From a deterministic interpolator to a design-based individual spatial predictor

Let us consider a bounded spatial domain $\mathcal{D} \subset \mathbb{R}^2$ and a function $z(\mathbf{u})$, $\mathbf{u} \in \mathcal{D}$. Furthermore, let $L = \{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be the set of n available locations, each with geographical coordinates $\mathbf{u}_i = (x_i, y_i)$. According to this notation, a spatial interpolator is defined as a weighted sum of the observed values; whether it assigns the true value to the known locations, the interpolator is defined as exact. A class of exact interpolators that generalizes the proposal of Shepard (1968) is

$$\widehat{z}(\mathbf{u}) = \begin{cases} \sum\limits_{i \in L} w_i z(\mathbf{u}_i), & \text{if } \mathbf{u} \notin L; \\ z(\mathbf{u}), & \text{otherwise}, \end{cases} \tag{1}$$

where the weights are standardized monotonically decreasing functions of the Euclidean distances between locations in the domain

$$w_i = \frac{f(\|\mathbf{u}_i - \mathbf{u}\|)}{\sum\limits_{i \in L} f(\|\mathbf{u}_i - \mathbf{u}\|)}.$$

The weighting system states, as a milestone of our future considerations, that geographical proximity favours the similarity of values of the variable under study.

Furthermore, let us suppose that we are interested in estimating the value of the variable under study at a finite number of locations in the domain $\mathcal{D}$ or that a (regular) finite grid has been superimposed over the spatial domain $\mathcal{D}$. In the former case, we assume that the coordinates are known for each location before sampling; in the latter, without loss of generality, we assume to know the coordinates of the centroid $\mathbf{u}_i = (x_i, y_i)$ of each grid element. Both cases fit the idea of finite population inference where a population $\mathcal{P}$ of size N is considered. By considering the number of spatial locations as finite, we are able to define the $N \times N$ symmetric matrix $\mathbf{\Phi}$ collecting the appropriate values of the function of Euclidean distances of each location from any other in the off-diagonal elements and having null diagonal values

$$\mathbf{\Phi} = \begin{bmatrix} 0 & f(\|\mathbf{u}_2 - \mathbf{u}_1\|) & \cdots & f(\|\mathbf{u}_N - \mathbf{u}_1\|) \\ f(\|\mathbf{u}_1 - \mathbf{u}_2\|) & 0 & \cdots & f(\|\mathbf{u}_N - \mathbf{u}_2\|) \\ \vdots & \vdots & \ddots & \vdots \\ f(\|\mathbf{u}_1 - \mathbf{u}_N\|) & f(\|\mathbf{u}_2 - \mathbf{u}_N\|) & \cdots & 0 \end{bmatrix}.$$

Moving to a finite population design-based framework, one can see the set L of known locations as the realization of a fixed-size probabilistic sampling design which can now be denoted as s, as customary in survey sampling. Formally, we define S as an arbitrary element of the σ-algebra of all possible samples belonging to the sample space; s is the realization of S. Let us suppose that the values $z(\mathbf{u}_1), \ldots, z(\mathbf{u}_N)$ refer to an unknown non-stochastic function, $z : \mathcal{D} \to \mathbb{R}$, evaluated at the sampled locations, $\mathbf{u}_1, \ldots, \mathbf{u}_n$. We propose a way to exploit the spatial information at the estimation level, stating that any sampling design without replacement appears more suitable because it guarantees that the spatial information will be used only once. Inclusion in sample S is therefore managed by the Bernoulli random variables $\mathcal{I}_{(i \in S)}$, $i = 1, \ldots, N$, such that $\pi_i = \Pr(i \in s) > 0$ for all $i \in \mathcal{P}$; the Bernoulli random variables can be collected in the N-dimensional random vector $\mathbf{q} = [\mathcal{I}_{(1 \in S)}, \ldots, \mathcal{I}_{(N \in S)}]^\top$. Analogously, inclusion of a generic t-tuple of locations $\mathbf{u}_i, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_{i+t-1}$, is managed by the Bernoulli random variables $\mathcal{I}_{(i,i+1,\ldots,i+t-1 \in S)}$ such that $\pi_{i,i+1,\ldots,i+t-1} = \Pr(i, i+1, \ldots, i+t-1 \in s) > 0$ for all $i, i+1, \ldots, i+t-1 \in \mathcal{P}$. Quantities $\pi_{i,i+1,\ldots,i+t-1}$ are the tth-order inclusion probabilities. Once a sample s is drawn from the population, the realization of the random vector $\mathbf{q}$ is the N-dimensional vector containing n unit values in correspondence of the sampled locations and zero otherwise. Exclusion from the sample is managed by the complement to one of the Bernoulli random variables managing inclusion, ($i.e.$ $1 - \mathcal{I}_{(i \in S)}$) because inclusion in and exclusion from the sample are two mutually exclusive events. These considerations lead to rewrite interpolator (1) as an individual design-based predictor as follows.

Let us rewrite the weighting system of interpolator (1) accordingly to the finite population design-based framework previously highlighted

$$\mathbf{h}_i = \mathcal{I}_{(i \in S)}\mathbf{e}_i + (1 - \mathcal{I}_{(i \in S)})\mathbf{q} \circ \boldsymbol{\phi}_i, \quad i = 1, \ldots, N, \tag{2}$$

where $\boldsymbol{\phi}_i = \boldsymbol{\Phi}\mathbf{e}_i$ is the ith column of matrix $\boldsymbol{\Phi}$, $\mathbf{e}_i$ is the ith N-dimensional canonical basis and $\circ$ is the (Hadamard) element-wise product. Then, the interpolator for the ith generic location can be rewritten as a design-based individual predictor, resulting in the ratio of linear combinations of random quantities

$$\begin{aligned}\hat{z}(\mathbf{u}_i) &= (\mathbf{h}_i^\top \mathbf{1}_N)^{-1}\mathbf{h}_i^\top \mathbf{z} \qquad (3)\\ &= \frac{\mathcal{I}_{(i \in S)} z(\mathbf{u}_i) + (1 - \mathcal{I}_{(i \in S)}) \sum_{j \neq i} \phi_{ij} z(\mathbf{u}_j) \mathcal{I}_{(j \in S)}}{\mathcal{I}_{(i \in S)} + (1 - \mathcal{I}_{(i \in S)}) \sum_{j \neq i} \phi_{ij} \mathcal{I}_{(j \in S)}}, \quad i = 1, \ldots, N,\end{aligned}$$

where $\phi_{ij} = f(\|\mathbf{u}_j - \mathbf{u}_i\|)$ is the ijth element of matrix $\boldsymbol{\Phi}$. Predictor (3) can be rewritten in matrix form to account for predictions at all the locations in the domain

$$\hat{\mathbf{z}} = \mathrm{diag}(\mathbf{H}\mathbf{1}_N)^{-1}\mathbf{H}\mathbf{z}, \tag{4}$$

where $\mathbf{z} = [z(\mathbf{u}_1), \ldots, z(\mathbf{u}_N)]^\top$ is the N-dimensional vector collecting the population values; and $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^\top$ is the $N \times N$ matrix collecting by row the non-normalized weighting vectors $\mathbf{h}_i$. Individual predictor (3), when the ith location is sampled, assumes the observed value; for unsampled locations, it computes expression (1), further adding uncertainty according to the principles of design-based finite population inference.

3. STATISTICAL PROPERTIES OF THE INDIVIDUAL SPATIAL PREDICTOR

Individual predictor (3) is a ratio of linear combinations of random quantities and, therefore, its properties can be analytically obtained only in approximated form.

Theorem 1 *The first-order Taylor expansion approximated expectation of predictor (3) is*

$$E[\hat{z}(\mathbf{u}_i)] \simeq \frac{(\pi_i \mathbf{e}_i + (\boldsymbol{\pi} - \widetilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{z}}{(\pi_i \mathbf{e}_i + (\boldsymbol{\pi} - \widetilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{1}_N}, \tag{5}$$

where vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$ *collects the first-order inclusion probabilities and vector* $\widetilde{\boldsymbol{\pi}}_i = (\pi_{1i}, \ldots, \pi_{(i-1)i}, \pi_i, \pi_{(i+1)i}, \ldots, \pi_{Ni})^\top$ *involves first- and second-order inclusion probabilities in the sample.*

Proof. See Appendix A.

Theorem 2 *The first-order Taylor expansion approximated variance of predictor (3) is*

$$V[\hat{z}(\mathbf{u}_i)] \simeq \mathbf{k}_i^\top E[\mathbf{h}_i \mathbf{h}_i^\top] \mathbf{k}_i \tag{6}$$

where we define vector

$$\mathbf{k}_i = \frac{(\pi_i \mathbf{e}_i + (\boldsymbol{\pi} - \widetilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{1}_N \, \mathbf{z} - (\pi_i \mathbf{e}_i + (\boldsymbol{\pi} - \widetilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{z} \, \mathbf{1}_N}{(\pi_i \mathbf{e}_i + (\boldsymbol{\pi} - \widetilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{1}_N)^2} \tag{7}$$

and the expectation of quantity $\mathbf{h}_i \mathbf{h}_i^\top$

$$E[\mathbf{h}_i \mathbf{h}_i^\top] = \pi_i \mathbf{e}_i \mathbf{e}_i^\top + (\widetilde{\boldsymbol{\Pi}} - \check{\boldsymbol{\Pi}}_i) \circ \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top, \tag{8}$$

which involves matrix $\widetilde{\mathbf{\Pi}}$ *collecting the column vectors* $\tilde{\boldsymbol{\pi}}_i$ *and matrix* $\breve{\mathbf{\Pi}}_i = E[\mathcal{I}_{(i \in S)} \mathbf{q}\mathbf{q}^\top]$*, which involve the inclusion probabilities up to the third-order.*

Proof. See Appendix A.

The individual predictor is finite population consistent (Särndal *et al.*, 1992)

$$\lim_{n \to N} \hat{z}(\mathbf{u}_i) = z(\mathbf{u}_i),$$

as interpolator (1) is exact.

Theorems 1 and 2 for simple random sampling without replacement (SRSWoR) are shown in Appendix B.

4. ASSESSMENT OF PREDICTOR PERFORMANCES

In this section, we evaluate and compare the performances of several individual predictors in varying spatial scenarios under different sampling designs. Our aim is to highlight situations where our proposal is competitive with respect to kriging, which is the benchmark model-based technique.

Design-based inference is rather rigid: it makes use of the spatial information by means of matrix $\mathbf{\Phi}$, which summarizes the inter-subjective shared geographical knowledge. A different aspect is represented by the choice of useful auxiliary information for varying probability sampling. Suitable auxiliary variables follow the basic geography laws, according to which individuals that are close tend to share more similar values than individuals that are far, as well as are strongly correlated with the study variable. We do not debate between design- and model-based inference in finite population, which has already been thoroughly examined, for instance, by Hansen *et al.* (1983), Little (2004) and Wang *et al.* (2012).

In order to investigate the properties of predictor (3), we first take into account a real dataset. Then, we conduct a simulation study for better managing different aspects of spatial populations. A Monte Carlo experiment is performed in both cases for assessing the behaviour of the predictor.

4.1. A case study

Environmental literature is rich of spatial datasets, very often analyzed via statistical techniques by data collectors. When spatial data are processed by non-statisticians, non-stochastic predictors are usually adopted without associating uncertainty to them. As statisticians, with (3), we make a proposal able to attach a measure of uncertainty to the prediction; therefore, we highlight the design-based spatial predictor performances by starting from a popular environmental dataset.

We refer to a dataset available in the `R CRAN`'s library `geoR` (Ribeiro Jr and Diggle, 2001), in this way by giving the possibility of testing further methodological proposal. The Kattegat dataset consists of 70 measures of salinity (Diggle and Lophaven, 2006) taken in the estuaries between Denmark and Sweden limited by the North Sea (see Figure 1). Water exchange between the Baltic Sea and the North Sea through Kattegat is closely coupled to the wind conditions. Salinity is given in psu (practical salinity unit). In the Kattegat salinity varies from approximately 30 psu in the northern part to approximately 20 psu in the south; a decreasing large scale component in this direction might be assumed. Values above 30 psu are termed as oceanic environment while freshwater is characterised by values less than 0.5. Distances between locations range from 0.65 to 212.15 kilometres.

[Figure 1 about here.]

The population size leads to a number of possible samples that is not computationally manageable; thus, a Monte Carlo experiment is performed. The samples of the Monte Carlo experiment constitute a randomly drawn subset of the sampling space having a computationally manageable dimension; eight sampling fractions, varying from 0.05 to 0.40 by 0.05, are considered. For each sampling fraction, 1000 random samples are drawn and for each of them predictor (3) is computed under two different choices of function f of the Euclidean distances. Firstly, we consider the inverse squared distance between a specified location $\mathbf{u}'$ and the generic location $\mathbf{u}$ in the spatial domain

$$\phi_{ij} = \|\mathbf{u}_j - \mathbf{u}_i\|^{-2}, \tag{9}$$

in this way randomizing, under a design-based framework, the inverse distance weighted (IDW) interpolator proposed by Shepard (1968). Secondly, we use a Gaussian kernel of the Euclidean distances between a specified location $\mathbf{u}'$ and any generic other $\mathbf{u}$ in the domain

$$\phi_{ij} = \exp\left\{-\frac{\|\mathbf{u}_j - \mathbf{u}_i\|^2}{2r^2}\right\}, \tag{10}$$

where r is the bandwidth, or radius, of the kernel. Along with predictor (3), the kriging predictor (Cressie, 1993) and the individual predictive form of the estimator of the total (Bolfarine and Zacks, 1992), which assigns the sample mean to the unsampled elements of the population

$$\hat{z}(\mathbf{u}_i) = \begin{cases} z(\mathbf{u}_i), & \text{if } \mathbf{u}_i \text{ has been sampled;} \\ \bar{z}, & \text{otherwise,} \end{cases} \tag{11}$$

are also computed. Results are summarised in terms of bias and RMSE in Figures 2 and 3, respectively.

Each figure contains the eight boxplots of predictors' bias or RMSE at the different sample sizes. Predictor (11) (denoted in both figures as "BZ") gives, in terms of bias, the worst performances for all sampling fractions. The result was expected since this predictor does not account for spatial information. For the other three predictors, the spatial information enters in the weights according to equations (9) and (10) for IDW and Gaussian kernel (G in both figures), as well as in the estimation of the variogram and trend for kriging. The median bias is very close to zero for all sampling fractions. In particular for IDW and G, it is almost always negative and ranges from -0.16 to 0.09 and from -0.08 to 0.04, respectively. An exception occurs for the smallest sampling fraction: the kriging's bias is always positive and ranges from 0.08 to 0.14. The bias for IDW decreases as the sampling fraction increases whereas it is almost stable for G. The boxplot sizes decrease as the sampling fraction increases, meaning that the biases tend to be similar for each predictor considered; since they are close to zero, the predictor can be considered as unbiased.

[Figure 2 about here.]

In terms of RMSE (Figure 3), except for the smallest sampling fraction, estimator (11) is always the worst; performances of the other three predictors are almost equivalent. In the first plot, kriging presents a very large RMSE that can be ascribed to the small dimension of the sample and the difficulty of estimating reliable variograms. For the other sampling fractions RMSE varies from 1.21 to 2.47 for IDW, from 1.02 to 2.06 for G and from 1.081 to 2.63 for kriging.

[Figure 3 about here.]

As a conclusion, proposal (3) suits for the Kattegat dataset, since its performances, both in terms of bias and RMSE, are equivalent to the kriging's, without suffering of the model-based solution pitfalls.

4.2. Simulated data

A simulation study is performed to thoroughly investigate the behaviour of predictor (3) in comparison with the design- and model-based results in the case of spatially varying populations (in the spirit of Baffetta *et al.*, 2009; Vagheggini, 2013; Bruno *et al.*, 2013). This simulation study is necessary to assess via computation the analytic properties of predictor (3). The unmanageable sampling space led us to perform a Monte Carlo study to estimate the performances of the predictor. The spatial domain is a regular grid of dimension 15 × 15 on the unit square (*i.e.*, a total of 225 point units). Data are generated according to Gaussian random fields with fixed mean, μ, and sill, σ^2, equal to 2 and 4, respectively; the nugget effect, η, is null. Various spatial correlation structures are taken into account via three different semivariogram models (*i.e.* exponential, Gaussian and a pure nugget) in order to appreciate the performances of predictor (3) in relation to the geographical distribution of the data. The chosen parameterization fixes the same effective range ($h = 0.7$) for the exponential and Gaussian semivariogram models; the pure nugget semivariogram imposes a null effective range as it is an uncorrelated spatial process.

For each of the three superpopulation models so obtained, we generate 100 populations. For each of these, three different auxiliary variables are generated over the same spatial domain imposing linear correlation with the study variable of $\rho = 0.35, 0.50, 0.85$. Each auxiliary variable is used to compute the first-order inclusion probabilities as the proportions of the auxiliary variable values over their total in the varying probability without replacement sampling (VPSWoR).

For each of the populations obtained, the Monte Carlo experiment consists of drawing 1000 random samples using the SRSWoR and VPSWoR designs at three different sampling fractions (*i.e.*, f = 0.05, 0.10 and 0.15). For each sample, the value of the study variable at unsampled locations is predicted by means of predictor (3) under two different choices of function f of the Euclidean distances: IDW and G corresponding to equations (9) and (10), respectively. The kriging predictor and the individual predictive form of the estimator of the total (11) are computed as well (in the case of SRSWOR $\bar{z}$ in (11) is the sample mean, whereas for VPSWoR it corresponds to the usual probability weighted estimator). The comparison is then performed via the bias and root mean squared error (RMSE) computed for each point of the domain over the 1000 Monte Carlo samples. The values of the bias and RMSE so obtained, are then averaged over the 100 populations generated from each superpopulation model, allowing to reduce the randomness induced by the random field. In the Monte Carlo experiment, we impose the same semivariogram model used for generating the corresponding populations in the kriging predictions, also the initial values used in the REML parameter estimation are set equal to those of the superpopulation model thus preventing the danger of misspecification; this simulation context puts kriging in the best case scenario.

In this simulation we address the following four topics. First, we check whether the use of Euclidean distances improves design-based individual spatial predictions. Secondly, we assess how the sample fraction influences the performances of the design-based individual predictor by itself or compared to the kriging predictor. Third, we compare the results obtained using SRSWoR and VPSWoR. Finally, we study whether the spatial dependence structure influences the performances of the design-based individual predictor in itself and in comparison with the kriging predictor. As mentioned above, all evaluations are based on the bias and RMSE averaged over the 100 generated populations, as shown in Figures 4-7. Each figure is composed of nine plots, where each row collects the results of one of the three superpopulations and each column shows one of the three sampling fractions. Each plot contains the boxplots of the overall bias (Figures 4 and 5) or RMSE (Figures 6 and 7) for all the strategies of sampling designs and predictors. Light boxplots refer to the results of the SRSWoR, while the dark boxplots refer to those of VPSWoR.

Figure 4 presents the boxplots of the overall bias of all 225 points in the domain obtained in the cases of SRSWoR (light boxplots) and VPSWoR (dark boxplots) when the auxiliary variable is generated with a correlation of $\rho = 0.35$ with the variable under study. First of all, when comparing results of predictor (11) and our proposal, it appears that the use of spatial information reduces bias. The overall median bias in case of SRSWoR is approximately null, whereas when using VPSWoR, overestimation is detected. This is less evident when the data come from a superpopulation model with a Gaussian semivariogram. Regardless of the sample fraction and the semivariogram structure, the biases of the IDW and G predictors are approximately equal to that of kriging.

[Figure 4 about here.]

Figure 5 compares SRSWoR and VPSWoR when the auxiliary variable is generated with a correlation of $\rho = 0.50$. The results for the SRSWoR case are equal to those of Figure 4 because the auxiliary variable does not affect this sampling plan. In this case, the Monte Carlo experiment shows that the bias of VPSWoR slightly increases regardless of the sampling fraction and the semivariogram model, except for (11), which reduces its bias. Since the results when $\rho = 0.85$ have the same behaviour we do not show the corresponding figure.

[Figure 5 about here.]

Figure 6 compares the RMSE of the four predictors in the presence of different semivariogram models and at different sampling fractions; again, the light boxplots refer to SRSWoR, while the dark boxplots to VPSWoR. The auxiliary variable is generated with $\rho = 0.35$. In this case, estimator (11) shows the poorest performance. When increasing the sampling fraction, the RMSE of all spatial predictors decreases, with higher evidence when data are generated imposing a Gaussian semivariogram. Negligible differences appear between the RMSE obtained for SRSWoR and VPSWoR; therefore, SRSWoR could be adopted without reducing performance. When the exponential semivariogram model is used to generate populations (*i.e.*, the upper panels), the spatial predictors present similar performances: the kriging predictor and the two different specification of predictor (3) are very similar for all sampling fractions. For this spatial superpopulation, which is usually characterized by smooth behaviour in space, the three predictors are almost equivalent in terms of RMSE. When data are generated under the Gaussian semivariogram model (*i.e.*, in the central panels), sampling fractions affect the performances of the spatial predictors by reducing their RMSE. At the smallest sampling fraction, they present very similar performances; as the sampling fraction increases, the kriging predictor tends to perform better. This is likely due to the spatial clusters that characterize populations generated from random fields with Gaussian semivariogram. In this simulation, this peculiarity seems to be better captured by kriging rather than by the rather simple functions of the Euclidean distances (9) and (10). The results on the pure nugget superpopulation model (*i.e.*, the lower panels) show that kriging has quite a lower RMSE than the other techniques, particularly at the lowest sample size. As the sampling fraction increases, the RMSE slightly decreases, but less so for kriging; at $f = 0.15$ the three predictors have nearly identical performances. For the pure nugget case, the improvement in considering Euclidean distances (*i.e.*, IDW and G) in a design-based setting is still appreciable but not as evident as for the other superpopulations with different semivariogram models. This result is reasonable because the pure nugget model corresponds to a spatially uncorrelated process where employing distances in predictions is theoretically irrelevant. At each sampling fraction, the RMSE of the individual spatial predictor is always lower for populations generated by imposing a Gaussian semivariogram rather than an exponential one to the random field.

[Figure 6 about here.]

Figure 7 collects the boxplots of the overall RMSE for the SRSWoR (light), which is the same as in Figure 6, and for the VPSWoR (dark) when the auxiliary variable is generated with a correlation of $\rho = 0.50$. Differences with the results obtained for $\rho = 0.35$ are almost negligible except for the pure nugget model at $f = 0.05$. In this case, choosing VPSWoR produces lower RMSE, regardless of the technique used. Surprisingly, for VPSWoR with the pure nugget superpopulation model, increasing the sampling fraction leads to higher RMSE. Generating the auxiliary variable with a correlation of $\rho = 0.85$ produces results similar to those obtained by imposing $\rho = 0.50$ which are not reported.

[Figure 7 about here.]

5. CONCLUSIONS

Design-based individual prediction constitutes an inter-subjective alternative to kriging, since the specification of a superpopulation model, in the case of kriging, is based on the researcher's conjectures. On the contrary, the definition of matrix $\mathbf{\Phi}$, which is a collection of functions of the Euclidean distances between locations, can be negotiated before sampling.

In this paper, individual design-based spatial predictor (3) is compared with kriging and predictor (11). The design-based individual predictor performs well when the geographical relationship among elements, defined through the decreasing function of the Euclidean distances collected in matrix $\mathbf{\Phi}$, mimics the spatial distribution of the study variable over the entire domain, both in the environmental example and in the simulation experiment. According to our knowledge in this field, the proposal of a spatial design-based individual predictor represents a novelty. For any sampling design, the comparison of this proposal with predictor (11) shows that using spatial information in estimation reduces bias and, therefore, RMSE.

In our simulation study, when assessing kriging predictions, we choose to model the spatial correlation by adopting the same semivariogram models used for generating populations. Thus, we somewhat prevented the misspecification problem by placing kriging in the most favourable scenario. Therefore, the validity of this proposal is strengthened by the biased experimental conditions imposed as favourable to kriging, which are far from real world applications.

The comparison with kriging needs a deeper discussion. First, if some spatial correlation is assumed, bias and RMSE of predictor (3) appear very similar to those of kriging. At small sample sizes, in some cases, our proposal has even better performances (e.g. in the application to a real dataset). Conversely, when data are generated according to an uncorrelated spatial process (*i.e.*, the pure nugget model), kriging always shows better performances regardless of the sampling fraction, because it produces reliable estimates of the sill parameter. Despite this, kriging does not substantially outperform our proposal. Second, it is known that kriging predictions are not robust when sample sizes are small; since the few available spatial lags are not enough to obtain good estimates of the parameters of the semivariogram model. In fact, at the lowest sample size, predictor (3) has approximately the same null bias as kriging and RMSE values that in some cases outperform kriging. Adopting VPSWoR does not affect RMSE, but unfortunately produces higher bias. Varying probability sampling is known as a very appropriate tool for improving finite population inference, when auxiliary variables are available for building unequal inclusion probabilities is available. However, geographical information is available for all population elements, and may improve inference even when simple random sampling is performed.

In general, improvements in inference due to the knowledge of a quantitative auxiliary variable producing varying inclusion probabilities is lower than the direct use of geographical information in design-based inference.

This work is the first step towards the development of inference on population summaries (*i.e.*, means or totals), able to account for spatial information available before sampling, in agreement with the idea of model-assisted design-based inference proposed by Cassel *et al.* (1976) and Särndal *et al.* (1992).

ACKNOWLEDGEMENTS

The research of the first two authors was funded by a FIRB 2012 grant (project no. RBFR12URQJ; title: Statistical modelling of environmental phenomena: pollution, meteorology, health and their interactions) for research projects by the Italian Ministry of Education, Universities and Research.

REFERENCES

Baffetta F, Fattorini L, Franceschi S, Corona P, 2009. Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sensing of Environment* **113**(3): 463–475.

Bolfarine H, Zacks S, 1992. *Prediction Theory for Finite Populations*. Springer–Verlag, New York.

Bruno F, Cocchi D, Vagheggini A, 2013. Finite population properties of individual predictors based on spatial patterns. *Environ Ecol Stat* **20**: 467–494.

Brus D, de Gruijter J, 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma* **80**: 1–44.

Cassel CM, Särndal CE, Wretman JH, 1976. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**(3): 615–620.

Cicchitelli G, Montanari G, 2012. Model-assisted estimation of a spatial population mean. *Int Stat Rev* **80**: 111–126.

Cordy C, 1993. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters* **18**: 353–362.

Cox D, Cox L, Ensor K, 1997. Spatial sampling and the environment: some issues and directions. *Environ Ecol Stat* **4**: 219–233.

Cressie N, 1993. *Statistics for Spatial Data*. Wiley, New York.

de Gruijter J, ter Braak C, 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Math Geol* **22**: 407–415.

Diggle P, Lophaven S, 2006. Bayesian geostatistical design. *Scandinavian Journal of Statistics* **33**(1): 53–64.

Gregoire T, 1998. Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research* **28**: 1429–1447.

Hansen MH, Madow WG, Tepping BJ, 1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association* **78**(384): 776–793.

Li J, Heap A, 2008. *A Review of Spatial Interpolation Methods for Environmental Scientists*. Geoscience Australia, Canberra.

Little RJ, 2004. To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association* **99**: 546–556.

Little RJA, 2014. Survey sampling: Past controversies, current orthodoxies, and future paradigms. In Lin

X, Genest C, Banks D, Molenberghs G, Scott D, Wang JL (eds.), *Past, Present and Future of Statistical Science, COPSS 50th Anniversary Volume*, chapter 37, CRC Press.

Margalho L, Menezes R, Sousa I, 2014. Assessing interpolation error for space–time monitoring data. *Stochastic Environmental Research and Risk Assessment* **28**: 1307–1321.

Mitas L, Mitasova H, 1999. Spatial interpolation. In Longley P, Goodchild M, DJ M, DW R (eds.), *Geographical Information Systems: Principles, Techniques, Management and Applications*, volume 1, chapter 34, Wiley, London.

Ribeiro Jr P, Diggle P, 2001. geoR: A package for geostatistical analysis. *R-NEWS* **1**: 14–18.

Särndal CE, Swensson B, Wretman J, 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Shepard D, 1968. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*.

Stehman S, 2000. Practical implications of design-based sampling inference for thematic map accuracy. *Remote Sens Environ* **75**: 35–45.

Stein M, 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York.

Stuart A, Ord K, 1987. *Kendall's Advanced Theory of Statistics: Distribution Theory*, volume 1. Arnold, London.

Vagheggini A, 2013. *Employing Distances in Design-based Employing Distances in Design-based Spatial Estimation*. Ph.D. thesis, Department of Statistical Sciences, University of Bologna.

Wang JF, Stein A, Gao BB, Ge Y, 2012. A review of spatial sampling. *Spatial Statistics* **2**: 1–14.

Webster R, Oliver M, 2007. *Geostatistics for Environmental Scientists*. John Wiley and Sons, Chichester.

APPENDIX

A. PROOFS

A.1. Approximated Expectation

Following Stuart and Ord (1987), the expectation of predictor (3) is obtained in approximate form using the first-order Taylor approximation because the individual predictor is the ratio of linear combinations of random quantities.

Proof of Theorem 1. Starting from equation (3) the first-order Taylor approximation of the expectation is

$$\mathrm{E}[\hat{z}(\mathbf{u}_i)] \simeq (\mathrm{E}[\mathbf{h}_i]^\top \mathbf{1}_N)^{-1} \mathrm{E}[\mathbf{h}_i]^\top \mathbf{z}.$$

Expectation of vector $\mathbf{h}_i$ is

$$\begin{aligned}\mathrm{E}[\mathbf{h}_i] &= \mathrm{E}[\mathcal{I}_{(i \in S)}]\mathbf{e}_i + \mathrm{E}[(1 - \mathcal{I}_{(i \in S)})\mathbf{q}] \circ \boldsymbol{\phi}_i \\ &= \pi_i \mathbf{e}_i + (\boldsymbol{\pi} \circ \boldsymbol{\phi}_i - \widetilde{\boldsymbol{\pi}}_i \circ \boldsymbol{\phi}_i) \\ &= \pi_i \mathbf{e}_i + (\boldsymbol{\pi} - \widetilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i,\end{aligned}$$

where $\mathrm{E}[\mathcal{I}_{(i \in S)}] = \pi_i$ is the first-order inclusion probability which can be collected in the vector $\mathrm{E}[\mathbf{q}] = \boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$ and vectors $\mathrm{E}[\mathcal{I}_{(i \in S)}\mathbf{q}] = \widetilde{\boldsymbol{\pi}}_i = (\pi_{1i}, \ldots, \pi_{(i-1)i}, \pi_i, \pi_{(i+1)i}, \ldots, \pi_{Ni})^\top$ are retrieved.

A.2. Approximated Variance

For the same reasons as for the expectation, the variance of the individual predictor is obtained in approximate form using the first-order Taylor approximation.

Proof of Theorem 2. The resulting first-order Taylor approximation of the variance is

$$\begin{aligned}\mathrm{V}[\hat{z}(\mathbf{u}_i)] &\simeq \frac{\mathrm{E}[(\mathbf{h}_i^\top \mathbf{z})^2]\,(\mathrm{E}[\mathbf{h}_i^\top]\mathbf{1}_N)^2}{(\mathrm{E}[\mathbf{h}_i^\top]\mathbf{1}_N)^4} - 2\frac{\mathrm{E}[\mathbf{h}_i^\top \mathbf{z} \mathbf{h}_i^\top \mathbf{1}_N]\,\mathrm{E}[\mathbf{h}_i^\top]\mathbf{z}\,\mathrm{E}[\mathbf{h}_i^\top]\mathbf{1}_N}{(\mathrm{E}[\mathbf{h}_i^\top]\mathbf{1}_N)^4} \\ &\quad + \frac{\mathrm{E}[(\mathbf{h}_i^\top \mathbf{1}_N)^2]\,(\mathrm{E}[\mathbf{h}_i^\top]\mathbf{z})^2}{(\mathrm{E}[\mathbf{h}_i^\top]\mathbf{1}_N)^4} \\ &= \frac{(\mathrm{E}[\mathbf{h}_i^\top]\mathbf{1}_N\,\mathbf{z}^\top - \mathrm{E}[\mathbf{h}_i^\top]\mathbf{z}\,\mathbf{1}_N^\top)\mathrm{E}[\mathbf{h}_i\mathbf{h}_i^\top](\mathrm{E}[\mathbf{h}_i^\top]\mathbf{1}_N\,\mathbf{z} - \mathrm{E}[\mathbf{h}_i^\top]\mathbf{z}\,\mathbf{1}_N)}{(\mathrm{E}[\mathbf{h}_i^\top]\mathbf{1}_N)^4} \\ &= \mathbf{k}_i^\top \mathrm{E}[\mathbf{h}_i\mathbf{h}_i^\top]\mathbf{k}_i, \end{aligned} \tag{A.1}$$

where the last equality is obtained via algebraic manipulation.

Vector $\mathbf{k}_i$ is defined as

$$\begin{aligned}\mathbf{k}_i &= \frac{\mathrm{E}[\mathbf{h}_i^\top]\mathbf{1}_N\,\mathbf{z} - \mathrm{E}[\mathbf{h}_i^\top]\mathbf{z}\,\mathbf{1}_N}{(\mathrm{E}[\mathbf{h}_i^\top]\mathbf{1}_N)^2} \\ &= \frac{(\pi_i\mathbf{e}_i + (\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{1}_N\,\mathbf{z} - (\pi_i\mathbf{e}_i + (\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{z}\,\mathbf{1}_N}{(\pi_i\mathbf{e}_i + (\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{1}_N)^2}.\end{aligned}$$

Expectation of matrix $\mathbf{h}_i\mathbf{h}_i^\top$ is

$$\begin{aligned}\mathrm{E}[\mathbf{h}_i\mathbf{h}_i^\top] &= \mathrm{E}[\mathcal{I}_{(i\in S)}\mathbf{e}_i\mathbf{e}_i^\top] + \mathrm{E}[(\mathbf{q} \circ \boldsymbol{\phi}_i)(\mathbf{q} \circ \boldsymbol{\phi}_i)^\top] - \mathrm{E}[\mathcal{I}_{(i\in S)}(\mathbf{q} \circ \boldsymbol{\phi}_i)(\mathbf{q} \circ \boldsymbol{\phi}_i)^\top] \\ &= \mathrm{E}[\mathcal{I}_{(i\in S)}\mathbf{e}_i\mathbf{e}_i^\top] + \mathrm{E}[\mathbf{q}\mathbf{q}^\top] \circ \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top - \mathrm{E}[\mathcal{I}_{(i\in S)}\mathbf{q}\mathbf{q}^\top] \circ \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \\ &= \pi_i\mathbf{e}_i\mathbf{e}_i^\top + \widetilde{\boldsymbol{\Pi}} \circ \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top - \breve{\boldsymbol{\Pi}}_i \circ \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \\ &= \pi_i\mathbf{e}_i\mathbf{e}_i^\top + (\widetilde{\boldsymbol{\Pi}} - \breve{\boldsymbol{\Pi}}_i) \circ \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top.\end{aligned}$$

In the previous equation we define matrix $\mathrm{E}[\mathbf{q}\mathbf{q}^\top] = \widetilde{\boldsymbol{\Pi}} = (\tilde{\boldsymbol{\pi}}_1, \dots, \tilde{\boldsymbol{\pi}}_N)^\top$ collecting the second-order inclusion probabilities in the off-diagonal elements and the first-order ones in the diagonal, and matrix

$$\mathrm{E}[\mathcal{I}_{(i\in S)}\mathbf{q}\mathbf{q}^\top] \circ \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top = \breve{\boldsymbol{\Pi}}_i = \begin{bmatrix} \pi_{1i} & \pi_{12i} & \cdots & \pi_{1i} & \cdots & \pi_{1Ni} \\ \pi_{21i} & \pi_{2i} & \cdots & \pi_{2i} & \cdots & \pi_{2Ni} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ \pi_{i1} & \pi_{i2} & \cdots & \pi_i & \cdots & \pi_{iN} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ \pi_{N1i} & \pi_{N2i} & \cdots & \pi_{Ni} & \cdots & \pi_{Ni} \end{bmatrix},$$

which collects the third-order inclusion probabilities in the off-diagonal elements, except for the ith column and row, each containing the second-order inclusion probabilities as well as the diagonal except iith element, which is the first-order inclusion probability π_i.

B. CASE OF SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT

B.1. Approximated expectation

In the case of SRSWoR the approximated expectation (5) of predictor (3) reduces to

$$\mathrm{E}[\hat{z}(\mathbf{u}_i)] \simeq \frac{z(\mathbf{u}_i) + at_{1i}}{1 + at_{2i}},$$

which contains the population quantities $t_{1i} = \boldsymbol{\phi}_i^\top \mathbf{z}$ and $t_{2i} = \boldsymbol{\phi}_i^\top \mathbf{1}_N$ and the constant

$$a = \frac{N - n}{N - 1}.$$

In the SRSWoR case the residuals of the first-order Taylor approximated expectation are of order $O(n^{-1})$.

B.2. Approximated variance

In the case of SRSWoR, quantities (7) and (8) involved in the approximated variance of predictor (3) reduce to

$$\mathbf{k}_i = \frac{N}{n} \frac{(1 + at_{2i})\mathbf{z} - (z(\mathbf{u}_i) + at_{1i})\mathbf{1}_N}{(1 + at_{2i})^2}$$

and

$$\mathrm{E}[\mathbf{h}_i \mathbf{h}_j^\top] = \frac{n}{N}(\mathbf{e}_i \mathbf{e}_i^\top + b\,\mathrm{diag}(\boldsymbol{\phi}_i)^2 + c\,\boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top),$$

where constants derived from inclusion probabilities up to the third-order are

$$b = \frac{N - n}{N - 1} \frac{N - n - 1}{N - 2} \quad c = \frac{N - n}{N - 1} \frac{N - n}{N - 2}.$$

In the SRSWoR case the residuals of the the first-order Taylor approximated variance are of order $O(n^{-2})$.

FIGURES

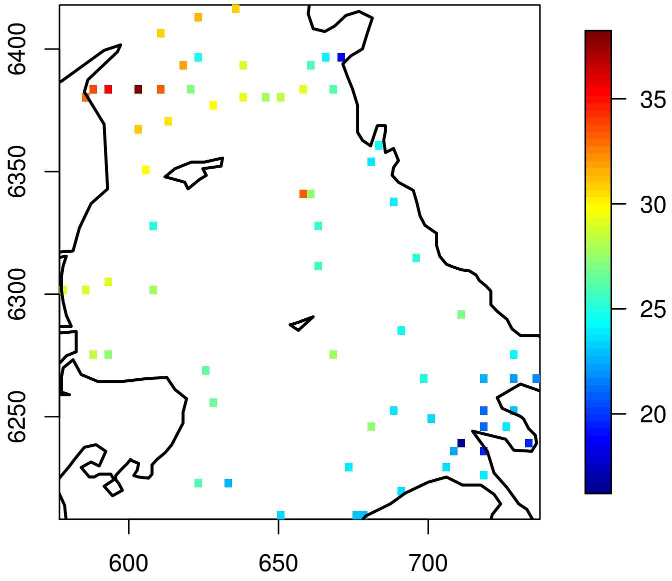

Figure 1. Locations and values of the Kattegat dataset.

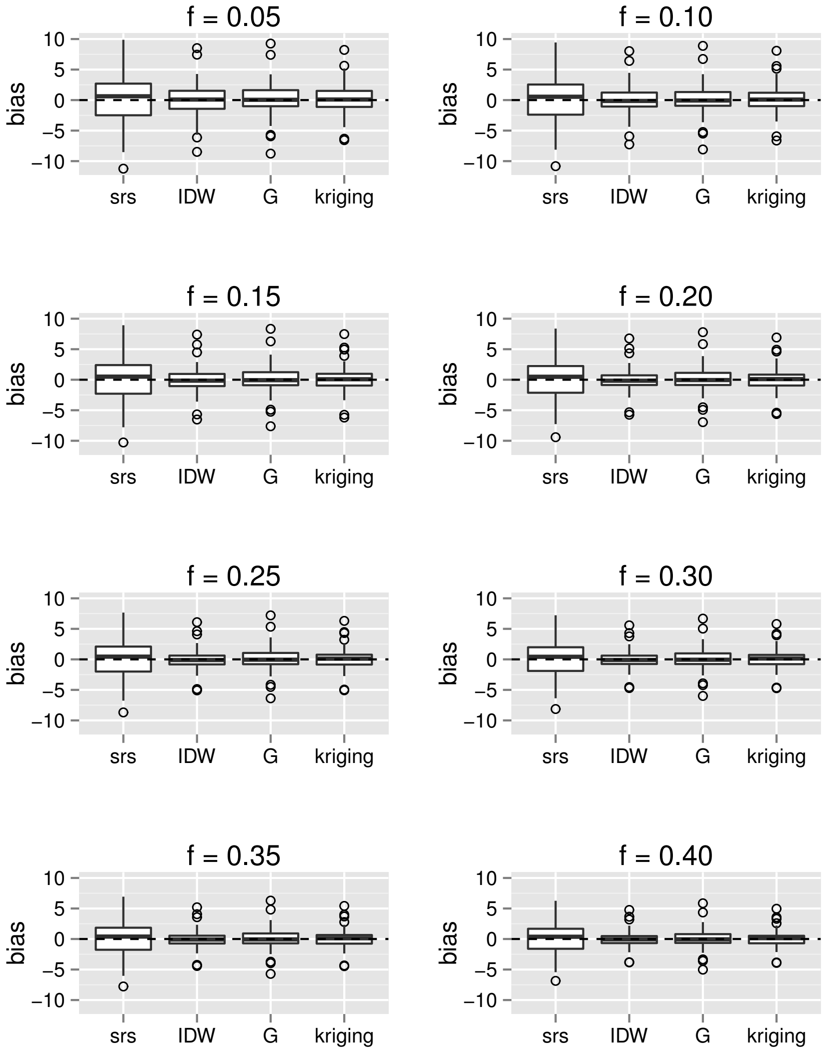

Figure 2. Bias for the Kattegat dataset.

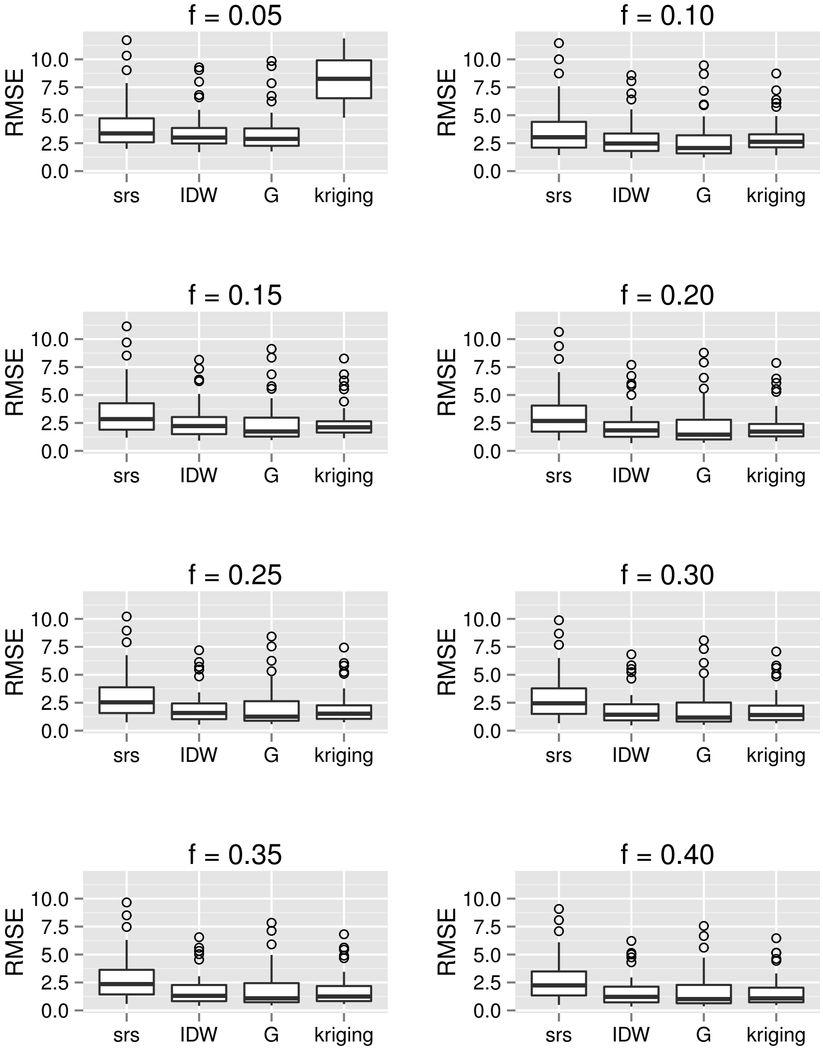

Figure 3. RMSE for the Kattegat dataset.

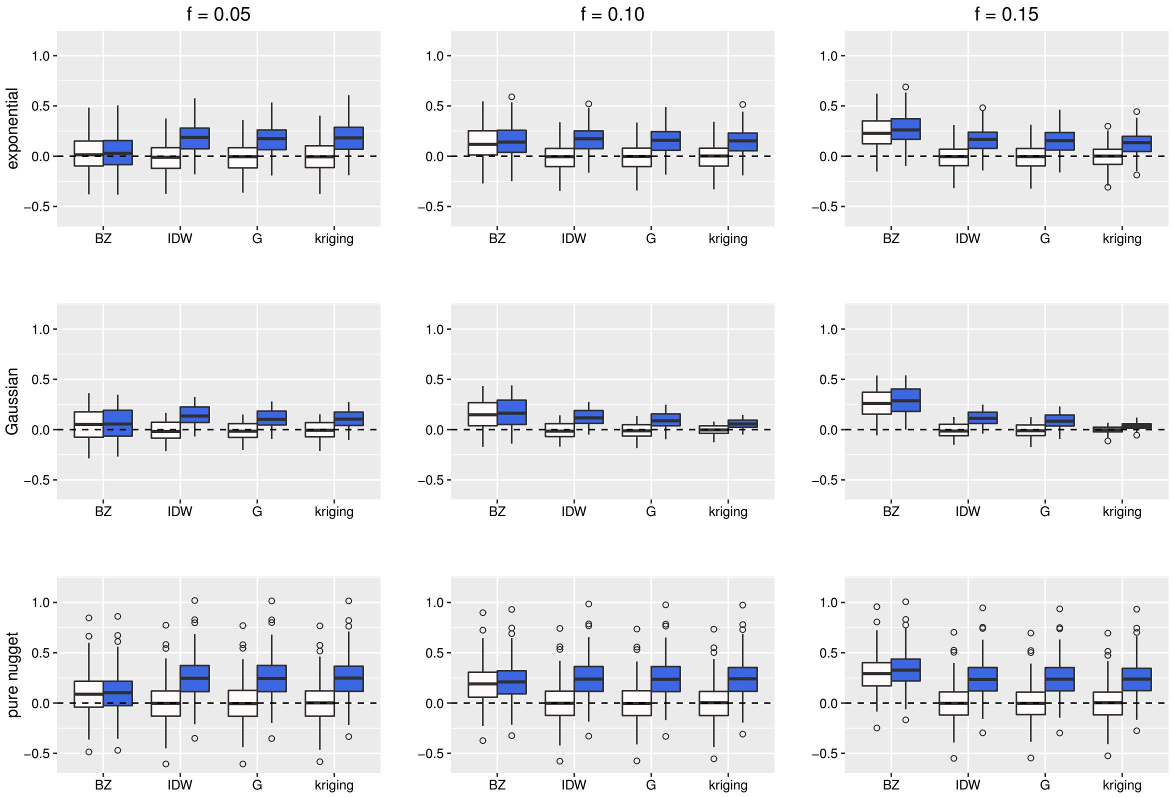

Figure 4. Overall bias; light boxplots are for SRSWoR and dark boxplots are for VPSWoR when the auxiliary variable has been generated with $\rho = 0.35$.

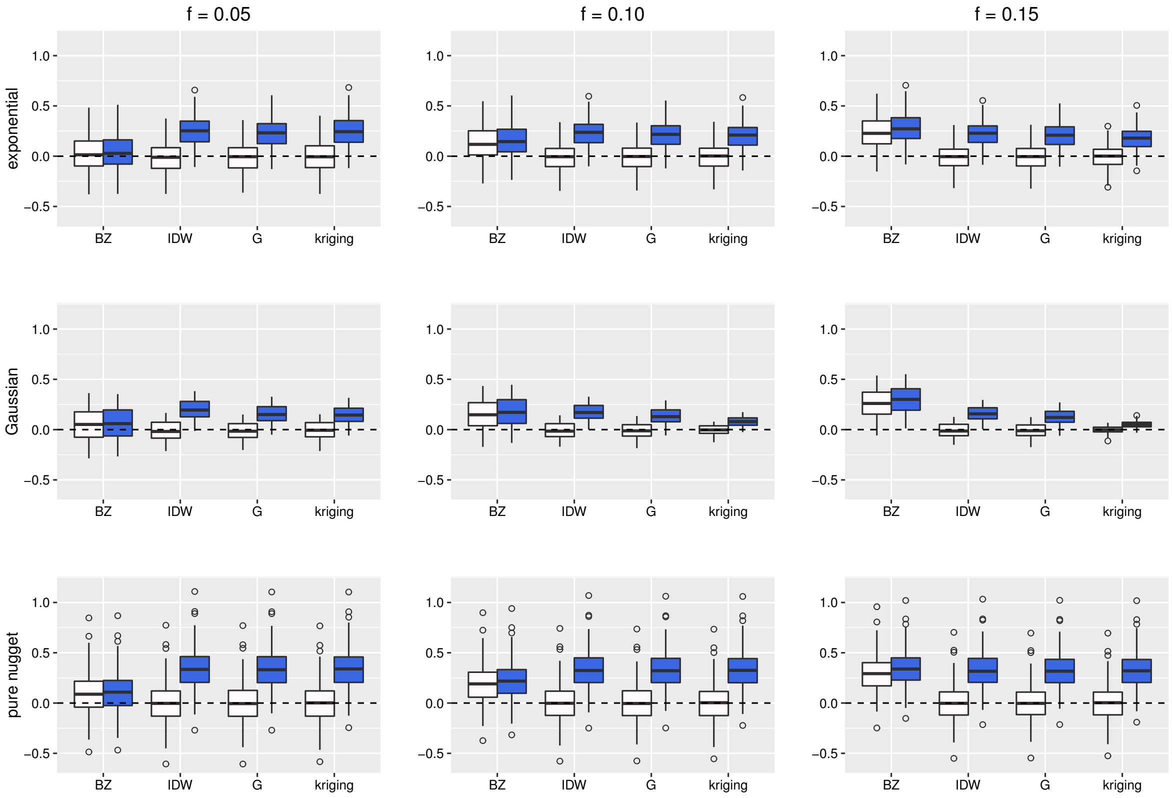

Figure 5. Overall bias; light boxplots are for SRSWoR and dark boxplots are for VPSWoR when the auxiliary variable has been generated with $\rho = 0.50$.

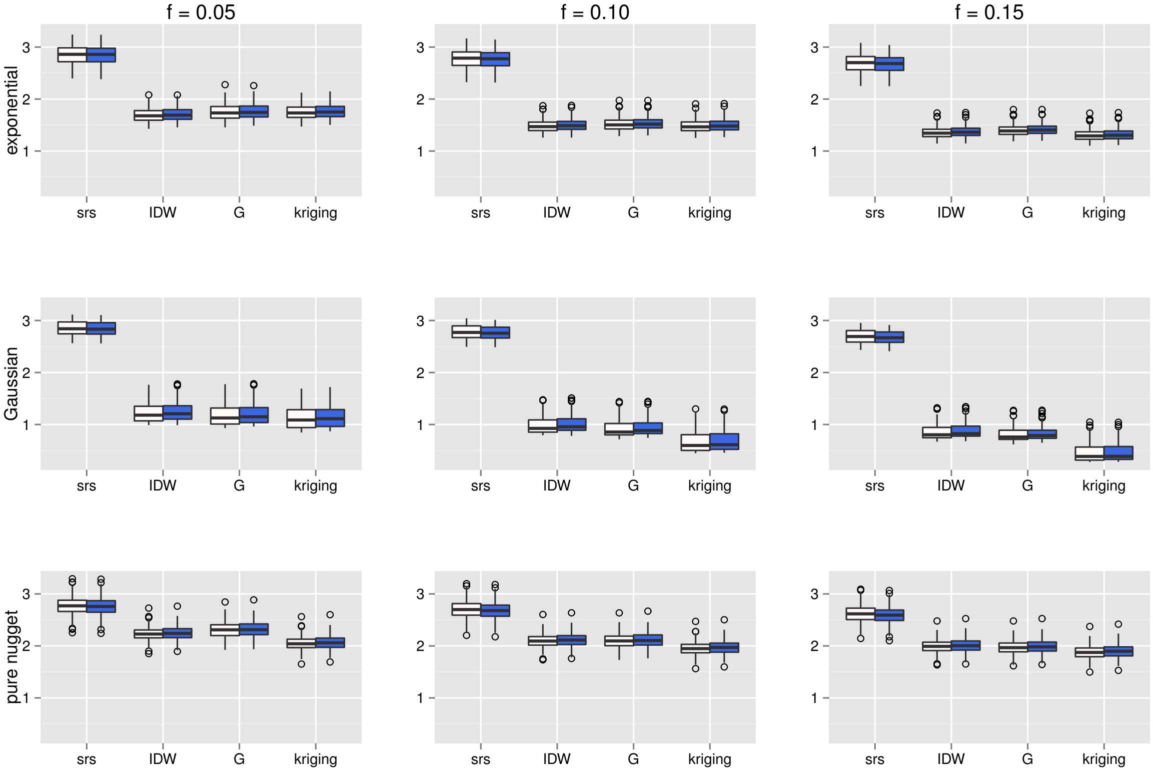

Figure 6. Overall RMSE; light boxplots are for SRSWoR and dark boxplots are for VPSWoR when the auxiliary variable has been generated with $\rho = 0.35$.

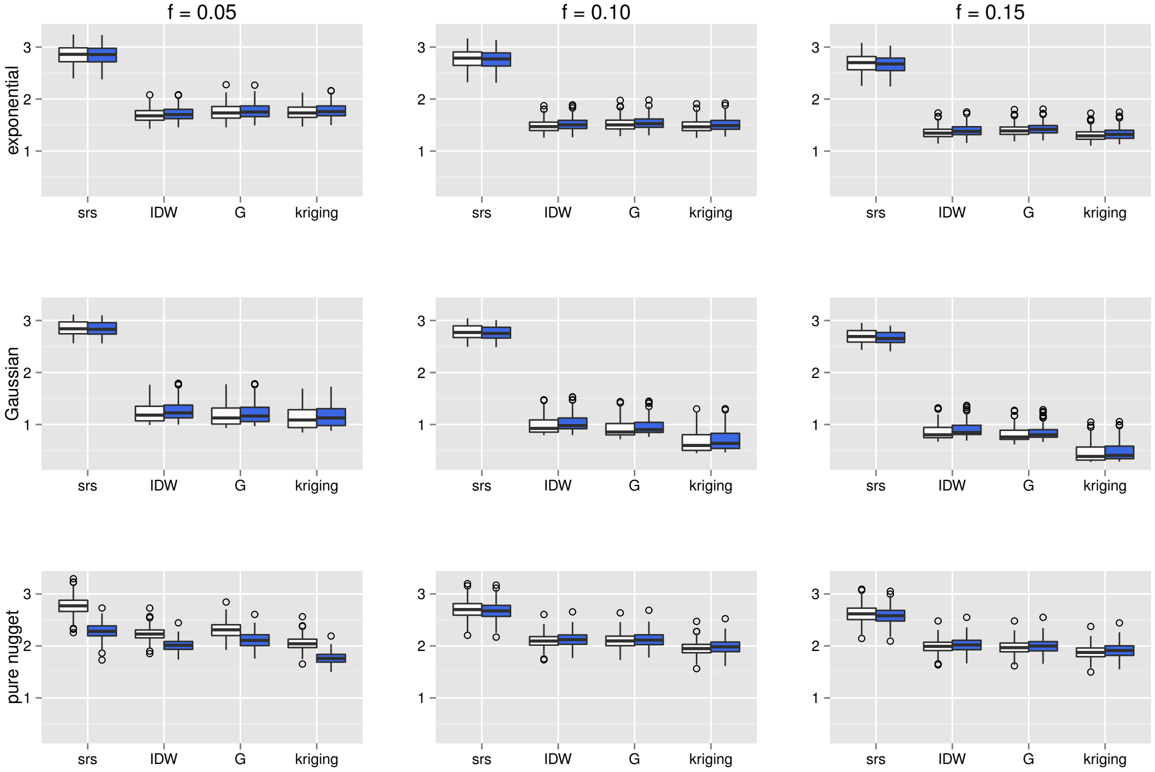

Figure 7. Overall RMSE; light boxplots are for SRSWoR and dark boxplots are for VPSWoR when the auxiliary variable has been generated with $\rho = 0.50$.