
# Individual design-based prediction: the assistance from spatial relationships

Daniela Cocchi[1]

[1] *Department of Statistical Sciences, University of Bologna; daniela.cocchi@unibo.it*

***Abstract.*** *Under the finite population design-based framework, spatial information regarding individuals of a population has traditionally been used for developing 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 a population characteristic available for each element before sampling. Individual predictors are obtained by reinterpreting deterministic interpolators under the finite population design-based framework, and their statistical properties can therefore be highlighted. We believe that this approach represents quite a novelty for spatial inference. Monte Carlo experiments help us to appreciate the performances of the proposed approach in comparison both with estimators that do not employ spatial information and with popular model-based proposals, i.e. kriging. We check whether the new predictor is suitable for inference and which are the most favourable conditions to its application. The performances of the achieved predictor are similar to those of kriging, especially for small sample sizes.*

***Keywords.*** *Design-based inference; Model-based inference; Spatial prediction; Spatial sampling*

## 1 Introduction

Prediction in finite populations is often performed under the model-based approach, where a superpopulation model is assumed. Samples are used for estimating model hyperparameters, which are at their turn the basis for the prediction of unsampled individual values [1]. Model misspecification is a real danger: the unsuitability of the proposed superpopulation model is a major criticism to model-based inference in finite populations.
design-based inference has a very long and established tradition in this context; the use of auxiliary information for differentiating the population units with respect to their propensity to be extracted in the sample or for enriching estimators is the strongest contribution for the reduction of the sampling error with respect to simple random sampling. The improvements are measured in terms of the reduction in the variability measures of the proposed estimators and rely on the consideration of the population elements as unknown non-random quantities, while the source of randomness lies on the discrete distri-

bution that manages sample drawing [10]. Synthetic population quantities, like totals, means or ratios, are the quantities that are most commonly estimated under the design-based finite population approach; these estimators can be expressed in predictive form.
For spatially structured populations, individual probabilistic prediction is commonly performed, using statistical tools expressed via superpopulation models, corresponding to the well-known kriging methods, which propose a stochastic setting opposite to the typical deterministic interpolators that are very popular in geography. model-based inference may suffer of two main drawbacks: the above mentioned risk of model misspecification, associated with the need of controlling the variability of parameter estimates. The two difficulties are contrasted by the specification of relatively simple, and therefore robust, models, and by the use of relative big samples.
design-based methods have not been in turn considered as suitable for spatially structured populations. This has been essentially due to a conceptual misunderstanding first highlighted by [3], after which in the last decades, a reappraisal of these methods has been promoted [5, 13]. The random extraction of population elements, too quickly seen only in the simple random sampling version, has been considered as contrasting with the forms of dependence among the elements of the population which are peculiar of spatial statistics methods. This misunderstanding derives from mixing two features that are conceptually very different. Dependence, or interdependence, regards the values of the variable under study, since the same individuals of a population may be independent with respect to the values of one variable and very strongly dependent with respect to the values of another. The presence or lack of independence between population characteristics may be confirmed by the sample, even if the sampling error cannot be avoided for the impossibility of making a census of the variable under study.
On the other hand, geographical characteristics are shared by all population elements. Geographical relationships pertain to individuals and are not related to the variables under study [7], therefore they can be a good basis for design-based inference. If a geographic relationship is considered relevant and is retained for subsequent analysis, it holds for the whole population and also, of course, for the sampled units. Indeed, a probabilistic sampling design does not randomize the values of the variable under study [7], rather randomization occurs for spatial locations. In the cases where, at the population level, the distribution of the variable under study is related to the geographical relationships between the elements of the population, the introduction of geographical relationships as auxiliary variables in the structure of design-based estimates is able to sensibly improve the performance of the estimators of synthetic population quantities [4], but also the performance of individual predictors.
Design-based inference on spatial data can exploit the population spatial information for developing more effective sampling designs or for enriching currently existing estimators. First, not only simple random sampling designs can be proposed, but also those involving varying probability: spatial information can be considered under different viewpoints in design-based sampling strategies [16]. Moreover, inclusion in or exclusion from the sample for the locations close to those already sampled can be managed (e.g. in adaptive spatial sampling). Second, the reappraisal of the design-based spatial inference leads to some new techniques for inference on individual values. For instance, [3] propose an individual estimator, which assigns the same value to all points in each subarea corresponding to the mean of the values observed at sampled locations belonging to each of the subareas in the entire domain.

## 2 Reinterpreting deterministic interpolators as individual design-based predictors

Individual spatial prediction is routinely performed either under a statistical model-based framework, via kriging, or via deterministic interpolators in a non-stochastic framework. Both approaches present some

drawbacks. Indeed, kriging may not be suitable when the available sample size is small, so the sample variogram estimation is performed using few couples, while the underlying model itself might be misspecified. On the other hand, deterministic spatial interpolators fail to capture the random nature of the phenomenon under study since they are based exclusively on geometric properties; thus, it is impossible to assess any probabilistic property for them [17, 11]. Deterministic spatial interpolators use weights that depend on the spatial distances between the locations of the available set, which are assumed as known. Probabilistic structures can be associated to such interpolators, so probabilistic properties assessment and uncertainty measures can accompany the obtained predictions. Indeed, geography is a population characteristic which is not restricted to the available data set, i.e. the sample in statistical jargon. So, when the set of available locations is seen as the outcome of a probabilistic sampling design, any deterministic interpolator can be interpreted under randomization. This consideration fits with the finite population design-based inference paradigm, where the weighting system associated to any estimator is constructed for the whole population before sampling. Recently, [2] proposed a new method for design-based spatial inference, leading to a predictor able to use the spatial information known at the population level for obtaining individual values at any location. Moreover, the geographical information on a specific location is not valuable in itself, but has to be considered with respect to any other element of the population. This population information can be summarized in a matrix, which, at each row/column, contains the information on the relationship (distance-type functions) between each population location and all the others. Following this idea, we consider a population of $N$ elements, which can correspond, however, to any kind of grid, and a sample of size $n$. So, $N$ can be huge and $n$ tend to 0, describing a situation of small sampling fraction.

According to the finite population terminology, when a relationship exists between labels and population values of the characteristic under study, labels are defined informative [14]. When dealing with spatial data under a design-based framework, we consider spatial coordinates as a further information associated to each population element. The availability of spatial coordinates indicates what we define as spatial informativity, which occurs when a relationship exist between spatial locations and the values of the variable under study. This point can be related to the solutions proposed in the case of preferential sampling [6, 12, 8]. Spatial coordinates constitute a very special kind of information that can participate or not to inference: geographical proximity can help inference that needs however to be assessed. We stress the point of keeping spatial coordinates separate from labels, that maintain the role of population elements identifiers; in this way simple random sampling, which is the simplest randomisation of population, is allowed, as a special case of varying probability sampling, and does not contrast with spatial inference. In agreement with spatial inference tools, spatial information is also separate from auxiliary variables.

## 3 Finite population design-based solutions may compete with model-based ones

Design-based individual predictors can be compared with well know benchmark model-based proposals, like kriging. The design-based version of deterministic spatial interpolators is useful when the variable under study has, for the whole population, a spatial distribution that mimics the geographical relationship among elements. Euclidean distances are the simplest example of general use of geographical information in statistical inference. In [2, 15] the conditions under which the proposed individual predictor for unsampled locations is preferable are assessed via Monte Carlo simulations.

Since kriging depends on the estimated variogram, when the model on which it is based is not misspecified, inference based on kriging outperforms individual design-based predictors. The strength of our solution is the separation between the values of the variable under study from the weights to associate

to sample values. The advantages of varying probability sampling can be taken into account; the idea of varying probability sampling is, at present, related to an auxiliary variable whose relative importance influences the selection of the populations elements in the sample. Such variable is related to the variable under study and is not necessarily related to the strength of spatial relationships. We propose to prepare, before sampling, a set of weights, different for each population value. The weights that will be used depend on the relationship between the site to predict and the set of sampled locations but are defined, however, before sampling. We are aware that behind any design-based solution a model is always hidden. Our effort resides in checking the extent to which a design-based solution to a finite population problem is practicable and when resorting to an explicit model specification is unavoidable.

## References

[1] Bolfarine, H. and S. Zacks (1992). Prediction Theory for Finite Populations.*New York: Springer Verlag*

[2] Bruno, F., Cocchi, D., Vagheggini, A. (2013). Finite population properties of individual predictors based on spatial patterns. *Environmental and Ecological Statistics* **20**, 467–494.

[3] 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.

[4] Cicchitelli, G., Montanari, G. (2012). Model-assisted estimation of a spatial population mean. *International Statistical Review* **00**, 1–16.

[5] Cox, D., Cox, L., and Ensor, K. (1997). Spatial sampling and the environment: some issues and directions. *Environmental and Ecological Statistics* **4**, 219–233.

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

[6] Diggle, P., Menezes, R., Su, T. (2010). Geostatistical inference under preferential sampling (with discussion). *Journal of the Royal Statical Society Series C-Applied Statistics* **59**, 191–232.

[7] de Gruijter, J., ter Braak, C. (1990). Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology* **22**, 407–415.

[8] Gelfand, A., Sahu, S., Holland, D. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics* **23**, 565–578.

[9] Ghosh, S., Gelfand, A., Mølhave, T. (2012). Attaching uncertainty to deterministic spatial interpolations. *Statistical Methodology* **9**, 251–264.

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

[11] Li, J., Heap, A. (2008). *A Review of Spatial Interpolation Methods for Environmental Scientists*. Geoscience Australia. Canberra.

[12] Pati, D., Reich, B., Dunson, D. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika* **93**, 35–48.

[13] Stehman, S. (2000). Practical implications of design-based sampling inference for thematic map accuracy. *Remote Sensing of Environment* **75**, 35–45.

[14] Thompson, S., Seber, G. (1996). *Adaptive Sampling*. Wiley. New York.

[15] Vagheggini, A., Bruno, F., Cocchi D. (2014) Individual spatial prediction under the design-based framework. *Joint METMAVII and GRASPA14 Workshop Proceedings*

[16] Wang, J.F., Stein, A., Gao, B.B., Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics* **2**, 1–14.

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