Design-based diagnostics for *k*-NN estimators of forest resources[1]

F. Baffetta, P. Corona, and L. Fattorini

Abstract: The *k*-nearest neighbours (*k*-NN) method constitutes a possible approach to improve the precision of the Horvitz–Thompson estimator of a single interest variable using auxiliary information at the estimation stage. Improvements are likely to occur when the neighbouring structure in the space of auxiliary variables is similar to the neighbouring structure in the space of the survey variables. Populations suitable for *k*-NN can be identified via the scores of the first principal component computed on the variance–covariance matrix of auxiliary variables. If the first principal component explains a large portion of the whole variability, distances among scores provide good approximations of distances in the space of auxiliary variables in such a way that the effectiveness of *k*-NN can be assessed by plotting the first principal component scores versus the sampled values of each of the interest variables. Monotone relationships with high values of Spearman's correlation coefficients should denote effectiveness. Otherwise, when the first principal component explains small fractions of the total variation, an index that directly quantifies the similarity between the neighbouring structure in the space of interest and auxiliary variables is proposed. The validity of the proposed diagnostics is theoretically argued and empirically proven by a simulation study performed on a wide range of artificial and real populations.

Résumé : La méthode des *k* plus proches voisins (*k*-NN) pourrait améliorer la précision de l'estimateur de Horvitz–Thompson d'une variable d'intérêt unique en utilisant une information auxiliaire au stade de l'estimation. Les améliorations devraient survenir quand la structure des voisins dans l'espace des variables auxiliaires est semblable à la structure des voisins dans l'espace des variables inventoriées. Les populations qui se prêtent à la méthode des *k*-NN peuvent être identifiées par les scores de la première composante principale calculés à partir de la matrice des variances et covariances des variables auxiliaires. Si la première composante principale explique une grande proportion de la variabilité totale, la distance entre les scores fournit une bonne approximation de la distance dans l'espace des variables auxiliaires, si bien que l'efficacité de la méthode des *k*-NN peut être évaluée en mettant en relation les scores de la première composante principale et les valeurs échantillonnées de chaque variable d'intérêt. Des relations monotones associées à de fortes valeurs du coefficient de corrélation de Spearman devraient être une marque d'efficacité. Par ailleurs, lorsque la première composante principale explique une faible proportion de la variation totale, nous proposons un indice qui quantifie directement la similarité entre la structure des voisins dans l'espace d'intérêt et les variables auxiliaires. La validité des diagnostics proposés est justifiée théoriquement et prouvée empiriquement par une étude de simulation réalisée sur une large gamme de populations artificielles et réelles.

[Traduit par la Rédaction]

Introduction

Several techniques for assessing natural resources couple information achieved from remotely sensed imagery with information collected from ground surveys. Of these methodologies, the *k*-nearest neighbours (*k*-NN) technique is becoming increasingly popular (e.g., Franco-Lopez et al. 2001; McRoberts et al. 2002; Tomppo and Halme 2004, among others).

The data necessary to perform the *k*-NN estimation of forest attributes are the digital numbers from multispectral remotely sensed imagery, which play the role of auxiliary variables and must be known for all the pixels in the application area. However, the values of the interest variables (forest and environmental attributes) are only known for a sample of pixels and, in most situations, are calculated using forest inventory information. Once remotely sensed and ground data are obtained, the *k*-NN estimate of the interest variables for a specific pixel are obtained by a weighted average of the values of the interest variables in the nearest pixels (with respect to a fixed metric in the space of the auxiliary variables) among those included in the sample.

At least to our knowledge, the only advanced investigations on the statistical properties of *k*-NN estimators have been performed by Kim and Tomppo (2006), McRoberts et al. (2007), and Magnussen et al. (2009) and lie in a complete model-based setting. Recently, Baffetta et al. (2009) derived the statistical properties of the *k*-NN estimators for

Received 12 January 2010. Accepted 24 August 2010. Published on the NRC Research Press Web site at cjfr.nrc.ca on 22 December 2010.

F. Baffetta[2] and L. Fattorini. Dipartimento di Metodi Quantitativi, Università di Siena, Piazza S. Francesco 8, Siena 53100, Italy.
P. Corona. Dipartimento di Scienze dell'Ambiente Forestale e delle sue Risorse, Università della Tuscia, Viterbo 01100, Italy.

[1]This article is one of a selection of papers from Extending Forest Inventory and Monitoring over Space and Time.
[2]Corresponding author (e-mail: baffetta@unisi.it).