STIMA CON METODO *k*-NN DI ATTRIBUTI FORESTALI SU SCALA REGIONALE: IL CASO DI STUDIO DELL'INVENTARIO FORESTALE DEL PIEMONTE

Nicola PULETTI (*), Roberta BERTINI (*), Gherardo CHIRICI (**), Fabio GIANNETTI (***), Davide TRAVAGLINI (*)

(*) geoLAB – Laboratorio di Geomatica, Dipartimento di Scienze e Tecnologie Ambientali Forestali, Università degli Studi di Firenze, Via S. Bonaventura 13, 50145 Firenze. Tel.: 055 3288621, Fax: 055 319179, davide.travaglini@unifi.it

(**) EcoGeoFor - Laboratorio di Ecologia e Geomatica Forestale, Dipartimento di Scienze e Tecnologie per l'Ambiente e il Territorio, Università del Molise, Contrada Fonte Lappone snc, 86090 Pesche (Is). Tel.: 0874 404113, Fax: 0874 404123, gherardo.chirici@unimol.it

(***) I.P.L.A. S.p.A., C.so Casale 476, 10132 Torino. Tel.: 011 8998933, Fax: 011 8989333, ipla@ipla.org

Riassunto

In questo lavoro viene descritta l'applicazione del metodo non parametrico *k-Nearest Neighbors* (*k*-NN) per la stima dell'area basimetrica di soprassuoli forestali a prevalenza di latifoglie. A questo scopo sono stati impiegati dati rilevati a terra nel corso dell'Inventario Forestale della Regione Piemonte, un'immagine telerilevata dal satellite Landsat 7 ETM+ e a altre informazioni ancillari come quota e pendenza del suolo derivate da un modello digitale del terreno. La migliore configurazione dell'algoritmo di stima è stata identificata con procedura di validazione *leave-one-out*. La valutazione dell'accuratezza delle stime prodotte dalla applicazione della migliore configurazione *k*-NN è stata applicata in 15 aree test di forma esagonale di raggio 5 km, confrontando la stima dell'area basimetrica ottenuta con metodo *k*-NN con quella derivata dalla elaborazione dei soli dati inventariali. I primi risultati ottenuti in termini di accuratezza delle stime prodotte sono positivi (errore quadratico medio percentuale pari all'11%) ed evidenziano le potenzialità del metodo *k*-NN per la stima di attributi forestali nel territorio investigato.

Abstract

This study describes the use of *k*-Nearest Neighbors (*k*-NN) method to estimate the basal area of broadleaves forests on a large area of Regione Piemonte on the basis of remotely sensed data and other ancillary information. Field data were acquired within the local Forest Inventory, satellite data were acquired by the Landsat 7 ETM+ multispectral bands and ancillary information (altitude and slope) were obtained by digital elevation model. The local optimal *k*-NN configuration was defined by a leave-one-out cross validation technique. The *k*-NN estimates were calculated for each pixel of the Landsat image inside 15 hexagonal test sites, each one of 5 km radius. Large area accuracy of *k*-NN estimates was calculated comparing the sum of the pixel level estimation by *k*-NN with the sum of the FI plots in each of the 15 test sites. Achieved results are satisfying (relative root mean square error of estimates of 11%) and they demonstrate the potential operative use of the *k*-NN method for forest attributes estimate over large areas.

Introduzione

Il metodo *k-Nearest Neighbors* (*k*-NN) è un metodo non parametrico che è frequentemente utilizzato per la stima di attributi qualitativi e quantitativi in campo forestale, esso coniuga informazioni acquisite nell'ambito di inventari su base campionaria con informazioni derivanti da telerilevamento (e/o da altre fonti informative ancillari georeferenziate). Il metodo *k*-NN è applicato

o è in fase di sperimentazione negli inventari forestali nazionali o regionali in USA, Finlandia, Norvegia, Svezia, Germania, Cina, Irlanda e Nuova Zelanda (Tomppo, 1991, Holmstrom et al., 2001; McInerney et al., 2005). Nel nostro Paese le applicazioni sperimentali di questo metodo sono piuttosto recenti (Maselli et al., 2005; Bertini et al., 2007).
Da un punto di vista concettuale il metodo *k*-NN permette di stimare il valore di una variabile Y per gli *N* elementi di una popolazione per i quali sia noto il valore vero di variabili ausiliarie (ancillari) correlate con Y, posto che per un campione di *n* elementi sia noto anche il valore vero di Y. In genere la popolazione è costituita dai pixel di un'immagine telerilevata multispettrale, la variabile Y è misurata a terra in corrispondenza degli *n* pixel del campione (detto *reference set*) e per tutti gli *N* pixel sono noti i valori di variabili ancillari rappresentate dai *digital number* (DN) delle singole bande spettrali, da indici ottenuti dalla combinazione di queste ultime e da altre eventuali informazioni correlate con i valori di Y (a esempio, quota, esposizione, tipo di suolo, ecc). Il valore incognito $\tilde{y}_j$ della variabile Y per ciascun *j*-esimo pixel dell'insieme *N-n* (detto *target set*) può essere stimato come media pesata dei valori di Y misurati in corrispondenza dei *k* pixels del *reference set* più vicini al *j*-esimo pixel nello spazio multidimensionale definito dalle variabili ancillari:

$$\tilde{y}_j = \frac{\sum_{i=1}^{k} w_{ij} y_i}{\sum_{i=1}^{k} w_{ij}} \quad [1]$$

dove il peso *w* può essere posto pari a *1/k* (in questo caso il valore $\tilde{y}_j$ è pari alla media aritmetica dei valori di Y misurati nei *k* pixel del *reference set* più vicini al j-esimo pixel) o, come avviene più frequentemente, può essere calcolato in modo inversamente proporzionale alla distanza multidimensionale tra il *j*-esimo pixel e ciascuno dei *k* pixel del *reference set* a esso più vicini (Corona, 2007).
La distanza multidimensionale può essere calcolata attraverso diversi tipi di misure, la più semplice delle quali è la distanza euclidea (De Maesschalck et al., 2000).
All'aumentare di *k* tende in genere ad aumentare l'accuratezza della stima $\tilde{y}_j$, ma con valori di *k* elevati si tende a ottenere una variabilità nei valori stimati per i singoli pixel minore di quella reale, a causa dell'effetto di livellamento prodotto dalla ponderazione.
La scelta delle variabili ancillari, del tipo di distanza multidimensionale e del valore di *k* è in genere condotta empiricamente attraverso una procedura *leave-one-out* (LOO) di valutazione dell'accuratezza delle stime prodotte. Questo tipo di procedura prevede la stima mediante *k*-NN del valore della variabile Y per ciascun *i*-esimo pixel del *reference set* con l'accortezza di escludere, ai fini della stima stessa, il valore vero y_i corrispondente a quel pixel: si ottengono così *n* valori stimati $\tilde{y}_i$ che confrontati con i corrispondenti valori veri y_i permettono di valutare l'accuratezza delle stime prodotte.
Sulla base dei risultati della procedura LOO viene definita la configurazione dell'algoritmo *k*-NN (in termini di variabili ancillari considerate, tipo di distanza multidimensionale, valore di *k*) che può fornire, nel caso indagato, le stime più accurate e che quindi viene applicata per la stima di $\tilde{y}_j$ sui pixel del *target set*.
Vari Autori (vedi ad esempio Reese et al., 2002) mettono però in evidenza che le stime prodotte con metodo *k*-NN sui singoli pixel dell'immagine sono caratterizzate, in genere, da livelli di accuratezza bassi e che gli errori si riducono se le stime fatte sui singoli pixel vengono utilizzate per derivare il valore dell'attributo incognito su porzioni di territorio più estese, mediando o sommando, a seconda dei casi, il valore stimato su più pixel. Si parla in tal senso di *pixel level estimation* e di *large area estimation*. Uno dei risultati attesi dall'applicazione di tali procedure di stima di variabili forestali ottenute da immagini telerilevate nell'approccio *large area estimation* è quello di individuare quale sia il livello di scala (ovvero l'estensione dell'ambito geografico di riferimento) a cui sia possibile

utilizzare le stime prodotte con *k*-NN rispetto a quanto sia possibile fare sulla base dei soli rilievi a terra inventariali.
In questo lavoro la valutazione delle accuratezze prodotte dal metodo *k*-NN nella stima dei valori di area basimetrica dei boschi a prevalenza di latifoglie nella Regione Piemonte è stata valuta con approccio *large area estimation* su aree test di forma esagonale di raggio 5 km (6495 ha) scelte in modo casuale. L'accuratezza delle stime è valutata confrontando la somma delle stime prodotte con *k*-NN per tutti i pixel a prevalenza di latifoglie con la soma dei valori misurati a terra nell'Inventario Forestale.

Materiali e metodi

Area di studio

La Regione Piemonte ha una superficie di 2.539.983 ha di cui 1.098.677 ricadono in aree montane, 769.848 in aree collinari e 671.458 in aree di pianura. Le superfici forestali, che ricoprono circa il 36% del territorio regionale, sono composte per il 24% da fustaie, prevalentemente lariceti, per il 62% da cedui, soprattutto di faggio e di castagno, e per il 14% da formazioni pioniere e d'invasione (Regione Piemonte, 2004). L'indagine è stata condotta nel settore centro-orientale della Regione Piemonte, nelle province di Torino, Alessandria, Asti, Biella, Cuneo e Vercelli. In questa zona i boschi di latifoglie sono costituiti per lo più da soprassuoli di castagno, faggio e querce, mentre tra i boschi di conifere prevalgono i larici-cembreti, le abetine e le peccete. Inoltre, sono presenti formazioni arbustive planiziali, collinari e montane, boscaglie pioniere e arbusteti subalpini (Regione Piemonte, 2004).

Dati a terra, dati telerilevati e informazioni ancillari

I dati di area basimetrica rilevati a terra provengono dall'Inventario Forestale della Regione Piemonte, realizzato in un arco temporale di quasi dieci anni, tra il 1996 e il 2004. I dati sono stati raccolti su aree di saggio circolari di raggio variabile (da 10 a 20 m) con centro posizionato sui vertici di una griglia semichilometrica coincidente col reticolato chilometrico UTM; la densità del campione inventariale è di un punto di campionamento ogni 25 ettari.
In questo studio sono stati utilizzati i punti inventariali che ricadevano nei boschi a prevalenza di latifoglie (Fig. 1) caratterizzati da valori di area basimetrica superiori a 10 m^2/ha; i punti dell'inventario che sulle immagini telerilevate sono coperti da nuvole o da loro ombre sono stati eliminati.

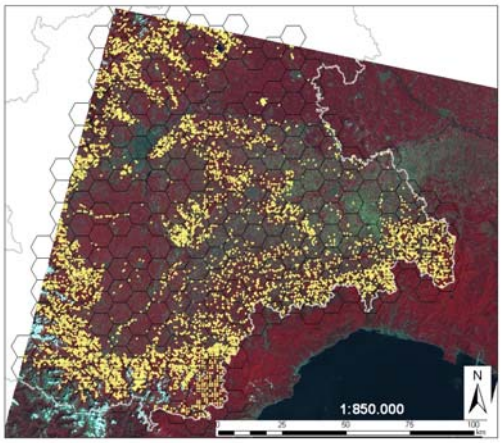


Figura 1 – Distribuzione geografica dei punti a bosco dell'Inventario Forestale della Regione Piemonte sulla base della scena Landsat 7 ETM+ utilizzata nello studio.

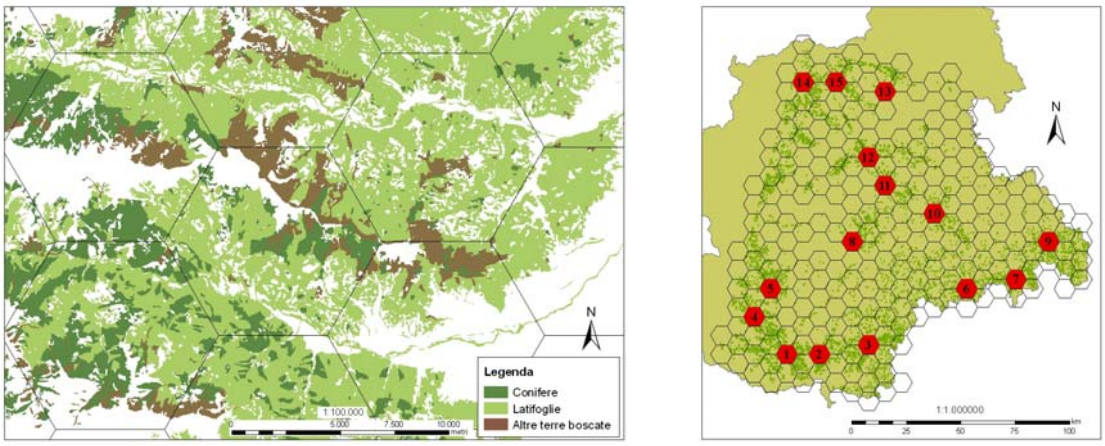
Complessivamente il *reference set* è composto da 5929 punti inventariali distribuiti su diverse tipologie di boschi (Tab. 1).

Le caratteristiche spettrali dell'area investigata sono state derivate dalle sei bande (esclusa quella termica) di una scena Landsat 7 ETM+ acquisita in data 23 luglio 2001. Altre variabili ancillari utilizzate nello studio sono la quota e la pendenza, calcolate a partire da un modello digitale del terreno con passo di 75 m.
La Carta delle tipologie forestali della Regione Piemonte, realizzata nel 2004 con scala di rappresentazione 1:25.000 (rilievo 1:10.000) (Regione Piemonte, 2004), è servita per definire la popolazione di riferimento, costituita dai boschi a prevalenza di latifoglie sui quali è stata eseguita la stima dell'area basimetrica con metodo *k*-NN.

Specie prevalente	*Punti inventariali*	
	n	*%*
Castagno	1862	30.7
Faggio	849	14.7
Querce varie	1251	20.5
Robinia	1234	21.9
Altre latifoglie	733	12.2
Totale	5929	100.0

Tabella 1 – Numero di punti inventariali (reference set) suddivisi per tipo di soprassuolo forestale

Configurazione sperimentale dell'algoritmo di stima e valutazione dell'accuratezza
In fase di ottimizzazione con LOO sono state testate diverse configurazioni del metodo *k*-NN per individuare la combinazione che in termini di *k*, metodo di calcolo della distanza multidimensionale, uso o meno di immagini satellitari normalizzate topograficamente (Puzzolo et al., 2005), era capace di produrre le migliori stime dell'area basimetrica sui pixels del *reference set*. In particolare, sono stati testati valori di *k* compresi tra 1 e 20 e cinque metodi di calcolo della distanza multidimensionale: euclidea (De Maesschalck et al., 2000), di Mahalanobis (Holmstrom et al., 2001), modificata con pesi fuzzy (Maselli, 2001), modificata con metodo regressivo multivariato e con pesi non-parametrici (Maselli et al., 2005).



(a) *(b)*

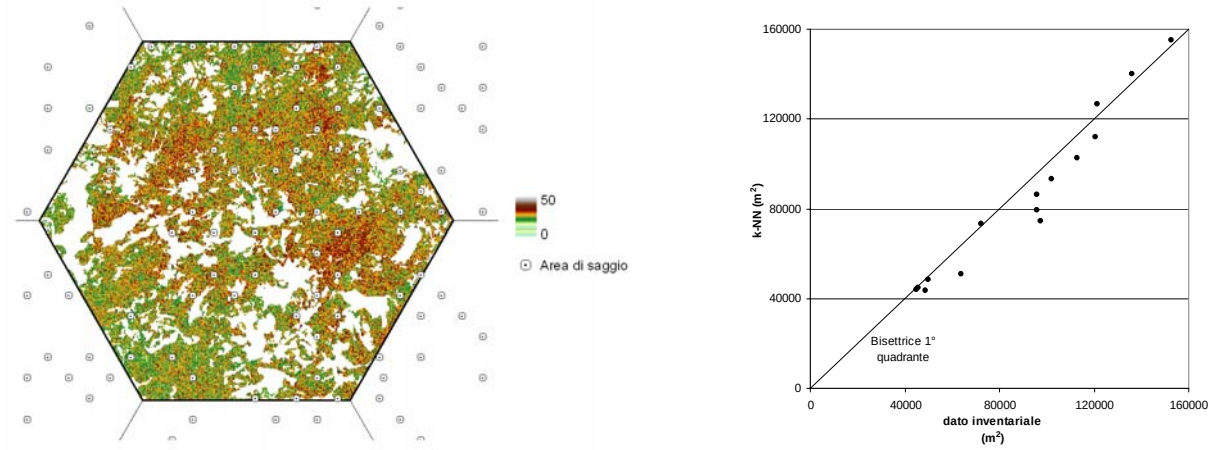
Figura 2 – a. Ingrandimento della carta delle tipologie forestali della Regione Piemonte. b. Suddivisione del territorio regionale in aree esagonali di raggio 5 km; in rosso i 15 esagoni utilizzati per valutare le accuratezze di stima del metodo k-NN

La combinazione ottimale individuata per l'area esaminata sulla base dei risultati del LOO, è stata applicata per stimare l'area basimetrica a ettaro dei soprassuoli forestali a prevalenza di latifoglie per ciascun pixel dell'immagine Landsat classificato a dominanza di latifoglie sulla carta delle tipologie forestali. All'interno dei 15 esagoni estratti casualmente è stata comparata la somma dei

valori totali di area basimetrica stimati con *k*-NN con la somma dei valori totali misurati a terra nelle relative aree di saggio dell'Inventario Forestale (Fig. 2). L'accuratezza delle stime prodotte è stata valutata calcolando lo scarto quadratico medio e lo scarto medio, ottenuti confrontando il valore di area basimetrica stimato con *k*-NN rispetto a quello misurato. Lo scarto quadratico medio percentuale e lo scarto medio percentuale sono stati calcolati come proporzione del valore medio delle stime (Mäkelä, Pekkarinen, 2004).

5. Risultati

Sulla base dei risultati ottenuti dal LOO, la migliore configurazione dell'algoritmo di stima è risultata quella che combina l'utilizzo di immagini satellitari normalizzate topograficamente con il metodo del *C-factor* (Puzzolo et al., 2005), con la distanza multidimensionale di Mahalanobis e un *k* pari a 8. Con questo tipo di combinazione lo scarto quadratico medio calcolato con LOO sui pixel del *reference set* è pari a 13,3 m^2/ha.
Successivamente tale configurazione è stata applicata per stimare l'area basimetrica a ettaro di tutti i pixel Landsat a prevalenza di latifoglie (Fig. 3a).
In Figura 3b per i 15 esagoni di riferimento è comparata la somma dei valori totali di area basimetrica stimati con *k*-NN per pixel con la somma dei valori totali dei punti inventariali. Nelle condizioni esaminate, le accuratezze delle stime *k*-NN su superfici esagonali di raggio 5 km sono caratterizzate da valori di scarto quadratico medio e di scarto medio di 9606 m^2 e 5701 m^2 rispettivamente; lo scarto quadratico medio percentuale e lo scarto medio percentuale sono risultati dell'11% e del 6% rispettivamente.



(a) *(b)*

Figura 3 – a. Esempio del risultato della stima k-NN dell'area basimetrica a ettaro in uno dei 15 esagoni per tutti i pixel Landsat a prevalenza di latifoglie. b. Confronto tra la stima k-*NN dell'area basimetrica nei 15 esagoni con il valore calcolato dai dati inventariali.*

6. Conclusioni

I risultati ottenuti in questa prima esperienza condotta in Piemonte, evidenziano le potenzialità del metodo *k*-NN per la stima dell'area basimetrica di soprassuoli boschivi a prevalenza di latifoglie, attraverso l'integrazione dei dati a terra dell'Inventario Forestale Regionale con immagini satellitari multispettrale Landsat 7 ETM+ e variabili ancillari derivate da un modello digitale del terreno.
Ulteriori indagini sono in corso per valutare l'affidabilità del metodo proposto ai fini della stima di attributi forestali su superfici di riferimento diverse da quelle testate, e per altre tipologie forestali.
Uno dei principali risultati attesi in questo tipo di studi è di individuare quale sia il massimo livello di scala a cui sia possibile utilizzare, tramite le stime da dati telerilevati con *k*-NN, i dati già acquisiti nell'ambito di un progetto inventariale su base campionaria. Il risultato ottimale atteso

dall'uso di queste metodologie sarebbe quello di poter utilizzare i dati inventariali anche per la stima in ambiti di gestione forestale di dettaglio, per esempio per la stima delle caratteristiche dei boschi a livello comunale o all'interno di un'area protetta (come quelle della rete Natura 2000), a supporto della redazione di piani di assestamento o per una prima valutazione dei danni subiti in un incendio forestale.
I metodi di stima come il *k*-NN diverrebbero quindi validi strumenti per massimizzare l'utilità (e il costo) degli inventari forestali e costituirebbero un valido legante tra le diverse scale della pianificazione e della gestione forestale.

Ringraziamenti
Il presente studio è stato realizzato nell'ambito di una collaborazione tra l'Istituto per le Piante da Legno e l'Ambiente (IPLA) e il Dipartimento di Scienze e Tecnologie Ambientali Forestali dell'Università degli Studi di Firenze.
Gli Autori ringraziano il Prof. Piermaria Corona dell'Università della Tuscia per la revisione del testo inerente la descrizione del metodo *k*-NN.

Bibliografia
Bertini R., Chirici G., Corona P., Travaglini D. (2007), *Confronto di metodi parametrici e non-parametrici per la spazializzazione della provvigione legnosa tramite integrazione di misure a terra, dati telerilevati e informazioni ancillari.* Forest@ 4 (1): 110-117.
Corona P. (2007). *Metodi di inventariazione delle masse e degli incrementi legnosi in assestamento forestale*. Aracne editrice, Roma.
De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L. (2000), *The Mahalanobis distance. Chemometrics and Intelligent Laboratori System*, 50:1-18.
Holmstrom H., Nilsson M., Ståhl G. (2001), *Simultaneus estimations of forest parameters using aerial photograph-interpreted data and the k nearest neighbor method*. Scand. J. For. Res., 16, 67-78.
Mäkelä H., Pekkarinen A. (2004), *Estimation of forest stand volume by Landsat TM imagery and stand-level field-inventory data.* Forest Ecology and Management, 196: 245-255.
Maselli F. (2001), *Extension of environmental parameters over the land surface by improved fuzzy classification of remotely sensed data*. International Journal of Remote Sensing, 17: 3597-3610.
Maselli F., Chirici G., Bottai L., Corona P., Marchetti M. (2005), *Estimation of Mediterranean forest attributes by the application of k-NN procedures to multitemporal Landsat ETM+ images*. International Journal of Remote Sensing, 17:3781-3796.
McInerney D., Pekkarinen A., Haakana M. (2005), *Combining landsat ETM+ with field data for Ireland's National Forest Inventory – a pilot study for County Clare*. Proceedings of ForestSat 2005, 31 May-June 3 2005, Borås, Sweden. 12-16.
Puzzolo V., Panizza M., De Natale F., Bruzzone L. (2005), *Correzione topografica di immagini Landsat TM e SPOT HRV in aree alpine orograficamente complesse*. Rivista Italiana di Telerilevamento, 32:67-77.
Reese H., Nilsson M, Sandström P., Olsson H. (2002), *Applications using estimates of forest parameters derived from satellite and forest inventory data*. Computers and Electronics in Agriculture, 37 (1-3):37-55.
Regione Piemonte (2004), *Tipi forestali del Piemonte - Metodologia e guida per l'identificazione*. A cura di IPLA S.p.a. - Istituto per le piante da legno e l'ambiente. Edizioni Blu, Torino.
Tomppo E. (1991), *Satellite image-based national forest inventory of Finland*. International Archives of Photogrammetry and Remote Sensing, 28(7-1):419-24.