Applied Mathematics and Computation 219 (2012) 2220–2230

Contents lists available at SciVerse ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Variable selection in regression models used to analyse Global Positioning System accuracy in forest environments

Celestino Ordóñez [a,*], Marta Sestelo [b], Javier Roca-Pardiñas [b], Enrique Covián [a]

[a] Department of Mining Exploitation and Prospecting, Polytechnic School of Mieres, University of Oviedo. Campus de Mieres, Gonzalo Gutiérrez, s/n, (33600) Mieres, Asturias, Spain
[b] Department of Statistics, University of Vigo. Campus Lagoas-Marcosende, (36310) Vigo, Pontevedra, Spain

ARTICLE INFO

Keywords:
Global Positioning System
Measurement accuracy
Forest canopy
Bootstrapping

ABSTRACT

Reliable information on the geographic location of individual points using GPS (Global Positioning System) receivers requires an unobstructed line of sight from the points to a minimum of four satellites. This is often difficult to achieve in forest environments, as trunks, branches and leaves can block the GPS signal. Forest canopy can be characterized by means of dasymetric parameters such as tree density and biomass volume, but it is important to know which parameters in particular have a bearing on the accuracy of GPS measurements. We analyzed the relative influence of forest canopy and GPS-signal-related variables on the accuracy of the GPS observations using a methodology based on linear regression models and bootstrapping and compared the results to those for a classical variable-selection method based on hypothesis testing. The results reveal that our methodology reduces the number of significant variables by approximately 50% and that both forestry and GPS-signal-related variables are significant.

1. Introduction

Knowledge of the shape and dimensions of the earth's surface and, more specifically, identification of the position of its characteristic features is a standard exercise in various branches of engineering, including forestry. The development of techniques based on the global navigation satellite and Global Positioning Systems (GNSS and GPS, respectively) have particularly transformed surveying practices [1]. The use of triangulation and traversing methods is no longer limited by the availability of a direct line of sight between a known position and the object whose position is to be determined [2].

Although satellite positioning systems provide reliable information on the position of individual points, irrespective of the weather conditions, at any time and place on or near the terrain surface, they require an unobstructed line of sight to a minimum of four satellites [3]. Use of GNSS techniques to study, describe and measure land used for forestry is an increasingly common practice. Studies of the application of GPS and GNSS techniques in forestry include plot inventories [4], cadastral surveys [5,6], map and plan making [7], geographic information systems [8], surface area and plot perimeter estimates [9] and even forestry planning and implementation [7]. However, tree cover reduces the effectiveness of these techniques due to the trunks, branches and foliage causing interference and signal loss [10]. This is evident in the lower precision—a difference of one order of magnitude—obtained regarding the position of characteristic terrain features [11]. Most studies

* Corresponding author.
E-mail addresses: ordonezcelestino@uniovi.es (C. Ordóñez), sestelo@uvigo.es (M. Sestelo), roca@uvigo.es (J. Roca-Pardiñas), covianenrique@uniovi.es (E. Covián).

0096-3003/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amc.2012.08.069

report a number of complications in using these techniques and provide practical recommendations for ensuring correct measurement.

The accuracy of GNSS observations, which needs to be consistent with the tolerance limits established for each case, depends on the systematic error component, mathematically expressed as bias, and the accidental error component directly related to precision (expressed as standard deviation). To ensure appropriately handle error and obtain sufficiently accurate measurements, it is important to determine possible causes of error during the observation phase and to assess their possible bearing on measurements.

Previous research has revealed that, along with conventional causes, several dasymetric parameters—tree dimensions, tree growth and standing volume—significantly influence accuracy in measurements made in forest environments. Bakula et al. [12], referring to real time kinematic observations, indicated the need to resolve ambiguities (a process called initialization) to ensure a high degree of precision and accuracy. Hasegawa et al. [13], who evaluated the accuracy of static-mode dual-frequency GPS receivers operating in forest environments, developed a model that estimates the probability of resolving ambiguities using logistic regression, with the observation period and tree cover index as independent variables, concluding that although position was more accurate when tree cover was less dense, 15 min of observation was sufficient to resolve ambiguities and obtain satisfactory precision under tree cover. Using a method based on genetic algorithms, Ordoñez et al. [14] concluded that dasymetric parameters had a greater bearing on positioning accuracy than variables associated with the GPS signal. Considering only the accuracy of vertical measurements, Wing and Frank [15] recorded significant differences between measurements made with GPS receivers with the same settings in environments with and without tree cover, concluding that forest cover had a negative influence on accuracy.

We describe a methodology for analyzing the relative importance of eight dasymetric parameters (arithmetic mean diameter, tree density, treetop height, Hart–Becking index, dominant height, basal area, standing volume and slenderness coefficient) and variables related to the GPS signal (signal-to-noise ratio in codes CA and P, position dilution of precision (PDOP), number of satellites transmitting signal, number of satellites receiving code and mean elevation angle) in the accuracy of GPS receiver observations made under tree cover. We used a linear regression model combined with bootstrap techniques to determine the minimum number of explanatory variables necessary to obtain the best prediction. To compare results, we used a stepwise backward method currently implemented in *R* program [16].

2. Methodology

For multiple regression models (with p variables), it is common to question what could be, and how to determine, the best subset of q ($q \leqslant p$) covariates that ensure the best possible fit to the data. This problem is particularly important in situations with many variables or with redundancy between highly correlated variables. In these contexts, the addition of a new variable to a model may appear to yield a better data fit, yet there are reasons why the estimates obtained may not be satisfactory: (a) the inclusion of irrelevant variables increases estimate variance and reduces model predictive capacity; and (b) the inclusion of a large number of variables complicates model interpretation. In our research we aimed to locate the variables with the greatest influence on the horizontal accuracy (H_{acc}) and vertical accuracy (V_{acc}) of GPS measurements under tree cover. We used an automatic forward stepwise method that selected the best subset of q variables that would ensure the best prediction capacity and eliminate the remaining variables, applying an optimization criterion based on the residual variance estimated by cross-validation. We then determined the minimum number of variables to be included in the model using the procedure described below.

Given a set of p initial variables, $X_1, X_2 \ldots, X_p$, for a given model of size k, where $\sigma^2(k)$ is the error variance obtained for the best subset of k variables, we have:

$$\sigma^2(k) = \min_{1 \leqslant j_1 < j_2 \ldots < j_k \leqslant j_p} \sigma^2_{j_1 j_2, \ldots j_k}, \tag{1}$$

where $\sigma^2_{j_1 j_2, \ldots j_k}$ is the residual of the model:

$$Y = \beta_0 + \beta_{j1} X_{j1} + \beta_{j2} X_{j2} + \ldots + \beta_{jk} X_{jk} + \varepsilon. \tag{2}$$

We used an automatic forward stepwise method that selected the subset of q variables that minimized the problem in Eq. (1). The procedure used was as follows:

Step 1. The covariate to be included in the first position is selected by fitting all possible models with one variable and then choosing the covariate which minimizes some error criterion (we used the residual variance estimated by cross-validation; see Eq. (6)). Similarly, the second covariate is selected by fitting all the possible models with two covariates fixing the previously selected covariate. The third covariate is introduced in a similar way but, in this case, the first two positions are established with the selected variables in the previous steps. This procedure is continued until the q covariates have been chosen. The procedure then starts again.

Step 2. For $j = 1, \ldots, q$, the jth actual selected covariate (keeping the remaining covariates) is replaced by another covariate to obtain a model with smaller residual variance.

Step 3. Step 2 is repeated until there is no change in the selected covariates. That is, the algorithm stops when a complete turn changes none of the q positions.