Selection of Tree-Size Variables for Appraisal Methods for Urban Trees According to Their Collinearity

Alvaro Sánchez-Medina, Esperanza Ayuga-Téllez, Lucrecia Contato-Carol, M. Ángeles Grande-Ortiz, and Concepción González-García

Abstract. The methods for appraising urban trees and municipal inventories in use today are expensive and require quantitative and qualitative variables with a high measurement cost. They are mathematically formulated from at least one tree-size variable to define a tree-size value. Researchers present a statistical methodology to analyze tree-size variables applied in appraisal methods for urban trees. A multivariate analysis method was carried out in order to obtain the lowest number of variables that explain the greatest variability of urban trees with no multicollinearity problems. The study was applied to urban trees in the City of Santiago del Estero, Argentina. The variables that showed the lowest collinearity were age and canopy area. The work includes a discussion of the use of correlated variables in appraisal methods for urban trees.

Key Words. Arboriculture; Multivariate Analysis; Tree-Size Value; Tree Appraisal; Tree Value.

The appraisal of urban trees is today a growing concern that affects many aspects of town planning. Increasing social awareness demands greater accuracy in the management of green infrastructures and urban trees. Policy makers are devoting more study and resources to this area, even in times of economic crisis.

Urban trees have traditionally been assigned a mainly ornamental function. However, they also play a recreational role and act as climate regulators, refrigerators (Skoulika et al. 2014), two other equally important functions that link man and nature and contribute to improving the environment and the welfare of its inhabitants (Tsunetsugu et al. 2013). Urban trees work as pollutant filters by reducing wind and thus preventing their dispersal, and also absorb acoustic stress, in addition to being a sign of good municipal planning (Ordóñez and Duinker 2013). Although these effects are widely recognized to increase property values and other benefits (Jiao and Liu 2010), it is difficult to set the price on the urban tree cover in the context of competitive markets (Contato-Carol et al. 2005), as their assigned market value is below their actual social value. This issue of environmental valuation is key to management and decision-making, and has been addressed in numerous studies (Campos et al. 2005; Redford and Adams 2009; Marinidou et al. 2013).

The lack of public woodland valuation is one of the main causes underlying its deterioration, and hinders its introduction in all areas of municipal management and territorial planning. Authors such as Fabbri (1989) and Caballer (1989), have noted the difficulties inherent in estimating the value of city trees, and caution that this estimation must be made according to the tree's utilities throughout its life span. Both authors also point to the conceptual difference between what is defined as the economic, environmental, and ornamental value of an urban tree. All these values represent the environmental and social services woodland contributes to society, and ultimately affect its economic worth, which must be established by considering a series of variables expressed in monetary terms.

The European Environment Agency states, “Green infrastructure is an important part of...
territorial identity and capital” and should be used to enhance the urban landscape. To preserve urban trees, municipal governments must thus conduct several long and costly procedures, such as surveys of residents’ willingness to pay to preserve and improve urban environments. Initiatives like the EU Inter-reg IVB project, Valuing Attractive Landscapes in the Urban Economy, implemented for northwest Europe, are designed to identify this willingness parameter by analyzing these types of surveys (EEA 2012).

One of the basic tools for the management of urban trees is the numerous appraisal standards and formulas existing worldwide. These all produce very different outcomes, as they use a variety of criteria and variables. The variables used in the assessment can be divided into three main groups: economic, state of health, and tree size. The most commonly used tree-size variables are canopy size, height, age, normal circumference, and basal area. The question is whether any of these variables is more appropriate for appraising the tree, and whether there are redundancies that can justify the use of any one as opposed to others. The answers to these questions can increase efficiency in cadastral inventories and directly affect the management of urban forestry by reducing economic costs, while allowing these criteria to be standardized throughout the different appraisal methods.

A tree’s value is commonly explained by functions of biometric variables. The variables most frequently used are basal area, normal circumference, height, dbh (diameter at breast high), normal area of the trunk, volume, canopy area, and age (Grande-Ortiz et al. 2012). Age is the most important variable used in non-parametric assessment methods; that is, capitalization and mixed methods (Ponce-Donoso et al. 2009; Grande-Ortiz et al. 2012). No measuring is required if the municipality records woodland planting data. Although more information can be obtained by considering a greater number of variables, it should be noted that these variables evolve over time and may be closely associated with chronological variables, such as age.

When defining inventory procedures, it must be taken into account that trees are complex organisms, and a full description of all physiological and environmental aspects is thus unrealistic (Constable and Friend 2000). The valuation of physiological complexity and woodland must be simplified without any significant loss of information that may affect the outcome of the assessment.

Some studies explore the variables that estimate tree value while minimizing the loss of information on physiological characteristics and tree growth. Jutras et al. (2009), for example, select 11 optimal quantitative variables, whereas Larsen and Kristoffersen (2002) and Yang et al. (2005) focus on the measurement of a single variable such as dbh, crown volume, current growth, or height. The problem is complex: a greater number of variables provides a more accurate picture of the evolution of the trees over time, but also implies an increase in the cost of inventories. In recent years, photogrammetric measurement methods have been used to obtain data on tree size and health, but most of this research has been conducted in the forestry context (Wulder and Seemann 2003; Stenberg et al. 2008; Cabrera et al. 2014). Although these techniques are currently expanding to the study of urban trees (Jensen et al. 2005), they are not yet widely available in municipal management, and consequently, most cities still use traditional inventory procedures.

The aim of this paper is to find variables with which to evaluate urban trees (in terms of their monetary value) to reduce inventory costs and the number of processes involved, using multivariate statistical methods that focus on inter-variable relationships.

This kind of analysis has been employed in previous studies (Savva et al. 2002; Heynen and Lindsey 2003; Turner et al. 2005) to describe and model complex relationships among multiple variables measured in the same population (Peña 2002). Multivariate inference techniques with multivariate descriptive techniques appear in Hammitt (2002), Jutras (2008), LaPaix and Freedman (2010), Grahn and Stigsdotter (2010) and Ayuga-Téllez et al. (2011). Several researchers have already highlighted the importance of reducing the number of variables in urban tree inventories (Jutras et al. 2009; Östberg et al. 2013). A recent study of the physiological state of urban trees considered many different kinds of biotic and abiotic variables, and concluded that quantitative variables are preferable (Jutras et al. 2009). A significant model, independent of the tree species, was obtained with combinations of
the following variables: dbh, annual dbh increment, crown diameter, canopy diameter increment, height and height increment, canopy diameter, canopy volume, and canopy volume increment.

The method presented in this work uses multivariate analysis techniques to determine the most appropriate variables for the basic value for the appraisal formulas. This is achieved by selecting the variables that explain the greatest variability but have no conflicts of multicollinearity.

STUDY AREA AND METHODS
This method was applied in the City of Santiago del Estero, Argentina, which has a wide variety of trees, shrub, and herbaceous species. This is due to factors such as the age of the site (the city was founded in 1553), its location between different phytogeographic regions, and immigration from Central and Southern Europe. Roic and Villaverde (1998) detected 226 different plant species: 132 trees, 73 shrubs, and 21 erect shrub climbers. Of the 132 tree species, only 15 belong to the local flora (dry Parque Chaqueño); the remainder are from other phytogeographic areas of the country or other continents.

Data were collected from various squares and streets in the city. Only broadleaf species (Tabebuia impetiginosa, Tipuana tipu, and Citrus ×aurantium) were studied, as conifers are underrepresented, and their use is limited almost exclusively to private gardens and farms. The trees selected were located in places where their age could be determined, and a total of 145 individuals were included in the study.

The method presented is based on statistical multivariate analysis and is structured in three steps to describe the sample and to identify the tree-size variables that explain the greatest variability with no problems of multicollinearity. It is therefore possible to determine which variables provide redundant information.

Descriptive Statistics
A statistical description was made of the variables measured (and the variables calculated from the functions of the observed variables) in urban trees from three study areas. Six variables were selected: four directly measured in the tree and two calculated from these first four.

The variables measured directly in the trees were:
- Normal circumference (c): The trunk perimeter measured in centimeters, perpendicular to the tree axis and measured at 1.30 m above ground level.
- Height (h): The distance between the base of the trunk and the upper end of the canopy, measured on its axis, in meters.
- Canopy diameter (cd): The canopy width, measured by the projection of its two ends in the field, in meters. As most of the canopies project in irregular shapes, the criterion was to take the major axis and its perpendicular. The average value is obtained as the arithmetic mean.
- Age (age): The number of years since seed germination (or the sprouting of cuttings for vegetative propagating species) until the time of the measurement.

The variables calculated from the variables measured directly in the tree were:
- Canopy area (ca), which follows cd measurements through the expression:
  \[ ca = \frac{p \times cd^2}{4}, \text{expressed in m}^2. \]  
- Basal area (g), which is the surface of the intersection of the trunk with a plane perpendicular to its longitudinal axis, measured at 1.30 m above ground level. These sections are characterized by their irregularity, and usually have an elliptical shape. In order to simplify the calculation, it is considered to be circular, and is calculated using the expression:
  \[ g = \frac{c^2}{4 \cdot \pi}, \text{expressed in m}^2. \]  

The statistical values obtained for the description were the number of trees sampled, the arithmetic mean, median, standard deviation, minimum and maximum sample value, and the coefficient of variation for each variable, in addition to the calculation of covariance and correlation coefficients, which measure the linear dependence between them (Ayuga-Téllez et al. 2013). This descriptive analysis was performed in Statgraphics 5.1 (Martín Fernández et al. 2001).
Relationship Between Variables
A combined analysis of variables was performed. In this step, the variables analyzed were average circumference (c), height (h), canopy area (ca), and age (age). Researchers calculated three correlation coefficients: Pearson, Spearman, and Partials.

The Pearson correlation coefficient (r) was used to quantify the linear relationship between these variables, and measured the degree of fit to a straight line for all the observations. The Pearson correlation coefficient is quite sensitive to outliers, so the information provided was completed by calculating the Spearman coefficient (Rs). To complete this information, researchers calculated the Partials correlation coefficient to identify the correlations between two variables without the influence of the rest.

Multivariate Analysis
With the same variables as in the previous step, researchers then worked with three multivariate analysis techniques: cluster analysis, multiple regression model, and canonical correlations. This achieved two objectives: it eliminated redundant variables and established the relationship between groups of variables.

To determine the most appropriate variables for assessing urban trees, a cluster analysis, also known as automatic, or unsupervised classification, was conducted to eliminate the redundant variables (Peña 2002). The classification algorithm used was the Ward method, which starts with an overall calculation of the heterogeneity of the group measured with the distances between variables; if the variables are continuous, the distances are expressed as:

\[ d_{jh} = \sqrt{\frac{1 - r_{jh}^2}{2}} \]

where \( d_{jh} \) is the distance between variable \( j \) and variable \( h \), and \( r_{jh} \) is the Pearson correlation coefficient for variables \( j \) and \( h \).

The variables that explain the most variability for the sample can then be established by means of a multiple regression model and canonical correlations (Peña 2002). The correlations between the sets of variables were analyzed to see if the amount of information obtained with them is comparable to the information obtained with the rest of the variables. Two linear combinations were obtained with the variables considered. The following statistics were used to study their importance: eigenvalues, canonical correlation, Wilks lambda (\( \lambda \)), and chi-square (\( \chi^2 \)).

RESULTS
As shown in the descriptive statistics (Table 1), the variation coefficient ranges from 13.48 for the sample can then be established by means of a multiple regression model and canonical correlations (Peña 2002). The correlations between the sets of variables were analyzed to see if the amount of information obtained with them is comparable to the information obtained with the rest of the variables. Two linear combinations were obtained with the variables considered. The following statistics were used to study their importance: eigenvalues, canonical correlation, Wilks lambda (\( \lambda \)), and chi-square (\( \chi^2 \)).

The combination of variables between \( h \) and \( ca \) therefore has the lowest coefficients, in both the linear and in Spearman’s correlations. This suggests that these two variables could jointly explain more variability than any other combination.

Table 1. Summary of the descriptive statistics for each variable.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Average</th>
<th>Median</th>
<th>S.D.</th>
<th>Min</th>
<th>Max</th>
<th>VC</th>
</tr>
</thead>
<tbody>
<tr>
<td>c (cm)</td>
<td>97.42</td>
<td>96.87</td>
<td>24.26</td>
<td>60.87</td>
<td>138.47</td>
<td>29.16</td>
</tr>
<tr>
<td>d (cm)</td>
<td>31.24</td>
<td>30.96</td>
<td>7.5</td>
<td>20.58</td>
<td>44.08</td>
<td>28.52</td>
</tr>
<tr>
<td>g (cm²)</td>
<td>1213.34</td>
<td>1155.46</td>
<td>492.69</td>
<td>619.47</td>
<td>2186.39</td>
<td>49.08</td>
</tr>
<tr>
<td>h (m)</td>
<td>9.41</td>
<td>7.54</td>
<td>1.12</td>
<td>5.98</td>
<td>13.6</td>
<td>13.48</td>
</tr>
<tr>
<td>da (m)</td>
<td>5.62</td>
<td>5.60</td>
<td>1.15</td>
<td>3.98</td>
<td>7.76</td>
<td>19.17</td>
</tr>
<tr>
<td>ca (m²)</td>
<td>33.67</td>
<td>31.50</td>
<td>14.47</td>
<td>17.62</td>
<td>63.65</td>
<td>37.22</td>
</tr>
</tbody>
</table>

Note: Average = arithmetic mean of the variable in individuals; Median = divided values of the variable in two equal-sized sets; S.D. = standard deviation; Min = the lowest sample value; Max = the highest sample value; and VC = variation coefficient.
The trees with estimated age and other tree-size variables were considered to observe the correlation between age and the other three variables, so there were not enough trees with the four measures. The correlation coefficients between age and c were: \( r = 0.7707; \ R_s = 0.8055; \ \rho_{12} = 0.4501 \). The correlation coefficients between age and h were: \( r = 0.7185; \ R_s = 0.7141; \ \rho_{12} = 0.2734 \). And the results for the correlation between age and ca were: \( r = 0.5179; \ R_s = 0.6450; \ \rho_{12} = -0.0281 \).

Among the tree-size variables measured, ca was very well correlated with c, and c with h (these latter two, in turn, measured the tree volume).

A multivariate clustering analysis was performed to reduce the number of variables (Figure 1). The tree-size variables form a homogeneous group (zero distance), while age is a heterogeneous variable relative to the rest.

A multiple regression model was made. A linear combination between age and tree-size variables produced a coefficient of determination of 79.02%, and a linear combination according to the equation:

\[
\text{Age} = 0.701774 \cdot c + 0.368309 \cdot h - 0.029501 \cdot ca
\]

Equation 4 indicates that age is more related to the circumference (c) than to the rest. The model coefficients are significant with a \( P \)-value < 0.1 in fulfilling the assumptions of residuals.

With the clustering variables, the homogeneity of height (h), canopy area (ca), and average circumference (c) were observed. The importance of each variable in relation to the rest is studied with the results of the canonical correlations (Table 2).

The equations for the variables for \( F_1 \) (significant canonical function) are:

\[
\begin{align*}
U_1 & = 0.71638 \cdot c + 0.33089 \cdot h \\
U_2 & = 0.42432 \cdot ca + 0.71205 \cdot age
\end{align*}
\]

The statistical values listed in Table 2 indicate that the correlations in the second linear combination can be considered invalid. The first linear combination between groups of variables is statistically significant (with 99% confidence) and has a canonical correlation of 0.847712.

**DISCUSSION**

A city’s trees are the critical component of the green infrastructure. A tree inventory is the systematic gathering of information on the urban forest and its organization into usable information for

![Figure 1. Hierarchical clustering dendrogram.](image-url)
tree management. It is important to define clearly who will use the inventory and who will collect the data, as this will determine the amount of resources needed to complete the project. A key objective of any community should be to maximize the benefits of trees and minimize the costs in achieving these benefits (Escobedo and Andreu 2015).

The most commonly used appraisal (parametric and mixed) methods use only one of the following tree-size variables to obtain a monetary value for urban trees: canopy size, height, age, normal circumference, and basal area. In fact, urban managers prefer quantitative parameters for urban tree inventories, they and consider a combination of tree-size variables (e.g., dbh, height, crown diameter, crown volume) as the simplest model for addressing the complexity of urban trees (Jutras et al. 2009). The formulaic expert method (FEM) alone uses a combination of some of these tree-size variables for the trunk and crown, rather than age, as reliable data were unavailable for heritage trees in Hong Kong, as is the case in most cities (Jim 2006). In contrast, capitalization methods use age as the only variable (Grande-Ortiz et al. 2012). This work therefore aims to establish which variables can explain the most variability with no problems of collinearity, in order to reduce costs in urban tree inventories.

The combined analysis of variables was conducted using average circumference (c), height (h), canopy area (ca), and age, as the calculated variables are related to them by the corresponding calculation functions. Age can be considered a basis for comparison independently of species (Hegedüs et al. 2011). The basal area (g) can be calculated with the average circumference (c), which is the variable employed in the American method of valuation (CTLA 1992; CTLA 2000), and is one of the most commonly used (Grande-Ortiz et al. 2012). Height (h) is one of the most widely-applied tree-size variables in urban tree inventories (Wood 1999; Martin et al. 2011; Moskal and Zheng 2012; Shrestha and Wynne 2012), and is also used indirectly in FEM and through corrective measures of basic value (Grande-Ortiz et al. 2012). Canopy area (ca) is another variable for calculating tree size that is easy to measure with current methods of photogrammetry (McRoberts and Tomppo 2007; Walton 2008; Abd-Elrahman et al. 2010; Millward and Sabir 2011); it is also a feature of the CONTATO and FEM valuation approaches (Grande-Ortiz et al. 2012) and can be used to determine the Location Index (Ayuga-Téllez et al. 2011).

The correlation coefficient results could offer the option of eliminating some variables to make the valuation. Normal circumference and basal area are totally correlated, and height is correlated with either of the other two coefficients up to 0.70. According to the literature (e.g., Brack and Wood 1998; Peper et al. 2001; Linsen et al. 2005), dbh, basal area, and tree height parameters can be used to predict growth or a tree's dimensions, including height, crown height and radius, leaf area, and so on. Some researchers (Martin et. al. 2011) present equations to predict crown width, with dbh as a dependent variable with R^2 results of between 0.91 and 0.94. Others use tree-size variables to estimate the age of urban trees (Quigley 2004) in cases with age readout errors of no more than +/- 15% (Łukaszkiewicz and Kosmala 2008).

Appraisal methods could be considered that eliminate this redundancy and use variables that are closely linked to tree-size but reflect the greater variability between them, according to the requirements of the general linear model (Caballer 1989; Ayuga-Téllez et al. 2013). For example, ca, g, and age could be sufficiently representative of tree form if correlations and locations are known for each species. The canopy area variable is directly related with the amount of shade provided by urban trees, which—according to the literature—is one of the most valuable benefits for the inhabitants (Lohr et al. 2004; Mell et al. 2013; Delgado-Bueno et al. 2013). Some studies of spatial distribution of urban trees select tree canopy cover as the dependent variable, as it can be related with socioeconomic variables, such as wealth, minority populations, educational attainment, and others (Heynen and Lindsey 2003). Age is a variable that is always included in all capitalization and mixed methods (Grande-Ortiz et al. 2012; Ponce-Donoso et al. 2013) and is of great importance in assessing the overall environmental benefits of urban trees (McPherson 2007; Kenney 2008). The basal area is only directly used in the TEDESCO appraisal method (Bernatzky 1978).

The high correlation between the different tree-size variables reveals a multicollinearity problem.
The canopy area \((ca)\) variable has the least correlation with \(age\), and has the considerable advantage of being measurable without the need for fieldwork, for example with remote sensing techniques (Cabrera et al. 2014) or combined with other analyses. Canopy size may show higher correlation values with age, so in many surveys the determination of age is easier (Hegedüs et al. 2011). Any tree-size measure will enable information to be collected on the benefits of a particular tree. Using cluster analysis and canonical correlation, the information from the observations can be sufficiently explained with a single tree-size variable. The combination of \(age\) with the least closely correlated variable \((ca)\) provides the greatest possible amount of information and represents the data set without any redundancies.

The cost of collecting data on individual trees is directly related to the amount of information obtained on each tree and the expertise of the data collector. Each piece of information collected incurs a cost in labor, data manipulation, and archiving; it is therefore critical to collect the optimal amount of information on each tree.

Tree inventories in streets and parks may include the large-scale collection of data, such as canopy cover, forest type, and condition, or examine the specific condition of individual trees. This wide variation in scale presents problems and opportunities in terms of the management level used in a particular community. For example, and related to outcomes of the present study, data collection using aerial photographs or by satellite now allows the analysis of the canopy surface variable with increasing reliability.

Multivariate analysis allows the variables that measure tree growth to be considered as similar and therefore requires the use of only one of the three variables to represent this group of measures. Age should be the variable to consider, along with one of the three tree-size measurements.

As there can also be said to be a strong dependence between both sets of variables for the data considered, age and one of the tree-size variables are sufficient to explain the information as a whole.

**CONCLUSIONS**

This work highlights the importance of maintaining records of both age and horizontal canopy area. Numerous urban tree inventories record the measurement of the canopy area, but only specify age qualitatively (by assigning the tree a code: young, mature, or old) rather than quantitatively, even in cases such as Madrid, Spain, or Santiago del Estero, Argentina, where the municipality itself uses appraisal formulas based on the tree age.

The goal of reducing costs in urban forest inventories is a matter of increasing concern to urban forest managers. The results of this research point to age \((age)\) and canopy area \((ca)\) as the minimum variables to be measured for the appraisal of trees in the city of Santiago del Estero, with the lowest loss of variability.

The use of tree-size variables and their annual increments in methods for appraising urban trees implicitly includes the tree age. However, when resolving a linear model for economic appraisal, the correlation between all these variables allows for further simplification of the expressions; as in this study, which uses variables dbh, overall height, basal area, canopy area, and age.

Age is a key variable for calculating economic value through capitalization formulas that express it as a fixed value. This aspect of the assessment is essential to allow authorities to set sanctions, as occurs in Santiago del Estero. One recommendation is to record the date of planting of urban trees to reduce data collection costs.

**LITERATURE CITED**


Alvaro Sánchez-Medina (corresponding author)
E.T.S.I. de Montes, Forestal, y Medioambiental
Ciudad Universitaria s/n, Madrid 28040
Spain

Esperanza Ayuga-Téllez
UPM
Madrid, Spain

Lucrecia Contato-Carol
Santiago del Estero, Argentina

M. Ángeles Grande-Ortiz
UPM
Ciudad Universitaria s/n, Madrid 28040
Spain

Concepción González-García
Madrid, Spain
Résumé. Les méthodes de relevés des arbres urbains et des inventaires municipaux utilisées actuellement sont onéreuses et nécessitent la mesure de variables quantitatives et qualitatives à des coûts élevés. Elles sont mathématiquement formulées à partir d’au moins une variable de la dimension des arbres pour définir une valeur correspondante. Les chercheurs présentent une méthodologie statistique pour analyser les variables de dimension des arbres appliquées aux méthodes d’estimation des arbres urbains. Une méthode d’analyse multivariée a été appliquée de manière à obtenir le plus petit nombre de variables pouvant exprimer la plus grande variabilité d’arbres urbains, sans problèmes de multicolinearité. Cette étude a été appliquée aux arbres urbains de la ville de Santiago del Estero, en Argentine. Les variables qui présentaient la colinéarité la plus basse étaient l’âge et la surface couverte par la canopée. La recherche intègre une discussion sur l’utilisation de variables correlées dans les méthodes d’estimation des arbres urbains.


Resumen. Los métodos para evaluar los árboles urbanos y los inventarios municipales en uso hoy en día son costosos y requieren variables cuantitativas y cualitativas con un alto costo de medición. Se formulan matemáticamente a partir de al menos una variable de tamaño para definir un valor de tamaño de árbol. Los investigadores presentan una metodología estadística para analizar las variables de tamaño de árbol aplicadas en los métodos de evaluación de árboles urbanos. Se realizó un método de análisis multivariado para obtener el menor número de variables que explican la mayor variabilidad de árboles urbanos sin problemas de multicolinealidad. El estudio se aplicó a árboles urbanos en la ciudad de Santiago del Es- tero, Argentina. Las variables que mostraron la menor colinealidad fueron la edad y el área de la copa. El trabajo incluye una discusión sobre el uso de variables correlacionadas en los métodos de evaluación de árboles urbanos.