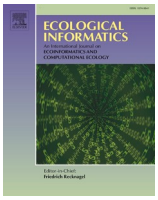
About polygon area uncertainty in GIS and its implications on agro-forestry estimates

Ecological Informatics 81 (2024) 102617

Contents lists available at ScienceDirect

Ecological Informatics

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



About polygon area uncertainty in GIS and its implications on agro-forestry estimates

Samuele De Petris *, Filippo Sarvia, Enrico Borgogno-Mondino

Department of agriculture, forest and food sciences, University of Torino, L.go Braccini 2, Grugliasco, Italy

ARTICLE INFO

Keywords:
GIS
Error propagation
Uncertainty
Area error
Shape metrics

ABSTRACT

Error affecting calculation of a polygon area from a digital map is an issue that is commonly neglected by remote sensing and Geographic information system (GIS) users. In this work, a method is presented aimed at estimating the uncertainty related to area calculation of polygons in a vector map. Additionally, area error relationship with polygon geometric features was analyzed, as well. A multivariate regression-based approach was applied for this task. After presenting the method, to demonstrate its operational capabilities, it was applied to 2 case studies corresponding (i) to a forest map and (ii) to the map from the Geo Spatial Aid Application used for in the framework of the EU Common Agriculture Policy. Estimated uncertainty (percentage) median values were found to be 0.01% and 0.02% for (i) and (ii), respectively. It was also demonstrated that the same polygon shape built using less vertices (longer vectors) generates area estimates that are less accurate than the ones from polygons built using a higher number of vertices (shorter vectors).

1. Introduction

In the framework of digital maps, especially when working at higher scales (e.g. cadastre), the area computation accuracy is an important issue to be considered. In fact, given a polygon defined by a sequence of vertices, its true "horizontal" area can be exactly known only if coordinates of vertices are assumed as "exact" (i.e. with infinite accuracy). This assumption is known to be always wrong, since national and international map standards rigorously define the expected horizontal accuracy of a map (σ_{xy}), once known its nominal scale (Goodchild et al., 2009; Hunter et al., 2000; Zhou and Stein, 2013). Uncertainty affecting the horizontal position of polygon vertices necessarily affects the accuracy of the corresponding computed area. In spite of this obvious consideration, quite often map users neglect this issue and assume as infinitely accurate area computation from digital maps by Geographical Information System (GIS) tools. This relies on an increasing unconsciousness about the true meaning of cartography and the rules it has to respect to be accepted as "official". Digitalization, geographical data sprawling provided through a huge number of geoportals online and a generalized lack of proper metadata are degrading consciousness of map utilization, especially from the accuracy issues point of view. The users, in fact, often use data and derive quantitative information from mapped features without having a complete knowledge of the data themselves nor o their quality (Leung et al., 2004). An unrealistic faith in geographical data accuracy seems to be the ordinary approach (Shi, 1998). Nevertheless, errors affecting length and area computation from geometric features of vector digital maps have been studied by many authors (Chun and Xiaohua, 2005; Crosetto et al., 2000; Rae et al., 2007). For example: points error distribution was analyzed in geodesy and surveying (Mikhail and Ackermann, 1976); uncertainty related to length computations was explored by Perkal (1956); while Chrisman and Yandell (1988) developed a formula to compute area variance; Kiiveri (1997) explored positional uncertainty in maps and Shi (1998) developed a statistical model for estimate errors of vectors in GIS.

According to the above mentioned scientific contributions, issues related to uncertainty affecting geometric features of digital maps are currently neglected while facing applications (Shi et al., 1999). Nevertheless, GIS applications are huge, and the demand for the resulting products is continually increasing especially in the agroforestry sector (Amici et al., 2017; Nowak et al., 2023; Sarvia et al., 2023).

This can be surprising if one considers some works that well demonstrate the operational impact of accuracy on agro-forest applications. For example, inaccuracies in delineating forest boundaries can lead to significant errors in estimating forest biomass (Næset, 1999). This misestimation can affect both ecological management and carbon stock assessments (Suwanlee et al., 2024) Additionally, the errors in area

* Corresponding author.
E-mail address: samuele.depetris@unito.it (S. De Petris).

https://doi.org/10.1016/j.ecoinf.2024.102617
Received 21 August 2023; Received in revised form 24 April 2024; Accepted 24 April 2024
Available online 27 April 2024

computation directly impacts economic valuation influencing land use planning, public taxation, and conservation funding (Judge and Allmendinger, 2011; van Oort et al., 2005).

Given these premises, uncertainty of geometric computation from vector features in GIS science, appears to be crucial to understand. Surprisingly, as already noticed by Chrisman and Yandell (1988), none of the GIS standard packages, is presently offering the capability of completing area computation with its expected accuracy. Some authors have implemented their procedures as external tools to be used within third-party software. For instance, Kiiveri (1997), developed a Geographic Resources Analysis Support System (GRASS) GIS (Neteler et al., 2012) add-on, and (van Oort et al., 2005)implemented an Environmental System Research Institute (ESRI) (Redlands, Ca) ArcInfo script. As far as authors know, the sole commercial software able to generate an estimate of the error affecting a polygon area is the Synergyse's TopoCheck® (Ljubljana, Slo). No standard and generalized approach seems, at the moment, to enter the most diffused and open GIS software (Heuvelink et al., 2006; Temme et al., 2009).

In this work, authors propose an operational approach implemented in QGIS open-source software and never proposed in literature, to easily face this challenge starting from a general rule that was at the basis of surveying and map production: every measure should be provided by two numbers, i.e. the value of the measure itself and the correspondent precision.

Estimation of the expected uncertainty of area computation can be obtained in three possible methods: (1) by comparison with another (more precise) dataset (Gross and Adler, 1996); (2) by statistical simulation (Caspary and Scheuring, 1992; Dutton, 1992; Shi, 1998) and (3) by error propagation analysis (Chrisman and Yandell, 1988). The first method requires a dataset of the same area to be used as reference. This should present a significantly higher level of accuracy, normally related to a higher map scale. This approach is not auto-consistent and it is therefore out of the scope of this paper.

The simulation approach provides results very similar to those obtainable by using error propagation (Shi, 2009), but has the disadvantage that cannot be described by a single formula; this makes it unsuitable to be easily implemented in a GIS (Shi, 2009), mainly due to the its computational efforts.

The last one considers area calculation as a problem of indirect measurement that can be faced applying the so called Variance Propagation Law (VPL) (Ku, 1966). In other words, area is intended as the result of a computation involving other direct measures. In this situation, an estimate of the expected area uncertainty can be obtained through the propagation of the errors affecting direct measures along the adopted formula. Direct measures involved in area computation are the coordinates of the polygon vertices. The starting point of such an approach is therefore to provide a measure of the uncertainty affecting coordinates of polygon vertices. This can be done only if the quality of the processed map is known through rigorous and reliable metadata that have to be necessarily coupled to the map itself.

The latter method was selected for this work due to its self-consistency, lack of reliance on external data, absence of statistical assumptions, and ease of integration into commonly used GIS platforms through ordinary geospatial algorithms.

After implementing the method, authors applied it to two case studies with the aim of analyzing the relationship between area uncertainty and polygon shape/size. The two selected case studies (NW Italy) refer to maps having different scale and content that provide the cartographic bases for two common agro-forestry applications. The first one is used for forest biomass computation; the latter is used for supporting payments within the Common Agriculture Policy (CAP). These maps are a solid benchmark since containing a wide variety of shapes, ranging from triangles to very complex polygons having thousands of vertices. The variety of shape and size of map features is an important issue to deal with. The dependence of area uncertainty from polygon shape/size was, in fact, already suggested in literature (Bondesson et al., 1998). To achieve this task authors propose five different geometrical indices useful for summarizing polygons shape.

2. Methods

2.1. Computing area of polygons in GIS

Polygon vector maps ordinarily entering GIS can be summarized as it follows: (i) maps obtained by digitalization of features from georeferenced image like the ones derived by remote sensing (Alvarez-Mendoza et al., 2019; Chow and Kar, 2017; Duarte et al., 2014; Gahegan and Ehlers, 2000); (ii) maps generated by automatic segmentation procedures and (iii) maps from ground surveys e.g., through global navigation satellite system (GNSS). Whatever the source of polygons, their area computation in GIS software is affected by the following approximations: (i) in general, it refers to the horizontal projection of the actual surface, while maps refer to its projection onto an ellipsoid; (ii) it is an under-estimation of the actual area since slope (local or average) is not considered; (iii) it is affected by an error that depends on the precision of the horizontal coordinates of the vertices defining polygon boundary.

To formalize the problem one has to consider that a *n*-sided flat polygon is made of *n* vertices (P_1, P_2, …, P_n) described by two coordinates $P_i = (x_i, y_i)$ connected by straight segments (vectors) in a closed loop ($P_1 = P_{n+1}$). The area is ordinarily computed in GIS by the so-called Gauss' formula (Eq. 1) where vertices are sorted counterclockwise (Zubaer et al., 2020).

$$A = 0.5 \sum x_i (y_{i+1} - y_{i-1}) \tag{1}$$

where x_i and y_i are the horizontal coordinates of the *i-th* vertex.

2.2. Accuracy of polygon vertex positioning

With reference to the above-mentioned problem, it is worth to stress that the proposed methodology can be only applied to bi-dimensional vector maps as implemented and managed by ordinary GIS software. Therefore, the work assumes the geometric approximation given by the vector layer (necessarily a sequence of straight segments) as the reference one. The number of vertexes, and consequently the polygon approximation of the mapped shape from the real world, depends on vector map providers and data quality check verified during the map validation stage. Shape approximation given by maps of real objects depends on the "nominal scale" the map was intended for when produced. Consequently, the number of vertices used to draw the map is assumed as satisfying approximation requirements imposed, natively, by the scale of the map. Consequently, only the horizontal positioning (and the related accuracy, σ_{xy}) is considered. This should be reported in the metadata of the maps or, eventually, deduced from the nominal scale the map is intended for. Since σ_{xy} is the ordinary accuracy measure supplied (or deduced) with maps, one has to go back to the disjointed values of errors affecting singularly *x* (σ_x) and *y* (σ_y) horizontal coordinates. Since no other assumption can be done, one has to admit that $\sigma_x = \sigma_y = \frac{\sigma xy}{\sqrt{2}}$.

Horizontal spatial accuracy of a map can be defined as the standard deviation of the errors affecting the horizontal coordinates, of a sample of check points from the map, whose *x* and *y* coordinates are compared with the correspondent ones from a reference source. This can correspond to ground surveyed points (by total station or GNSS), photogrammetrically derived points or, directly, maps showing an accuracy (or nominal scale) significantly higher than the one of the map to be evaluated.

Going back to polygon area computation, one can say that σ_{xy} depends on the nominal scale of the considered map. Often, digital maps are supplied together with metadata directly reporting σ_{xy}. Alternatively, it can be derived referring to national standards for map production that relate it to the nominal scale of the map.

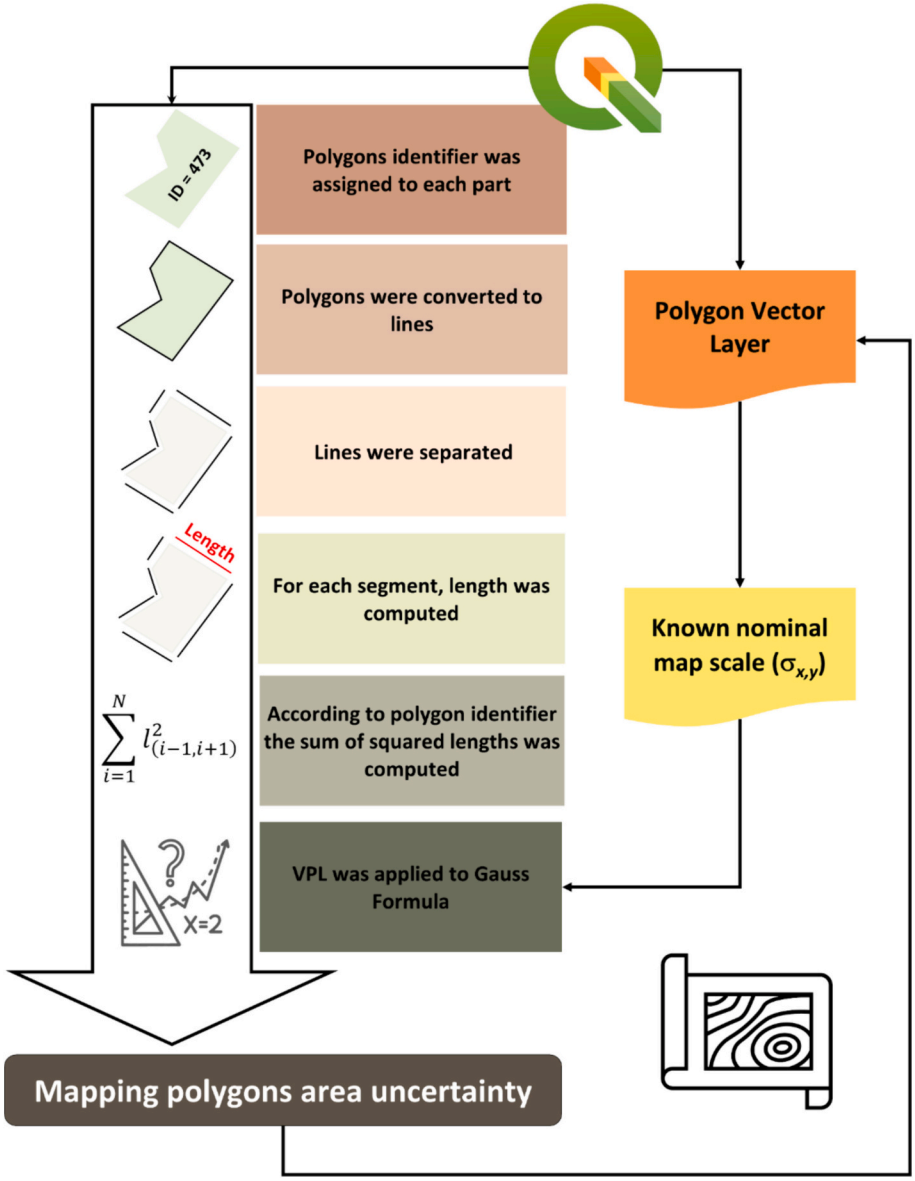


Fig. 1. Procedure implemented in QGIS to map polygons area uncertainty.

In this case, σ_{xy} can be obtained with regards to reference values defined in the drawing domain of maps (φ_{xy}), and corresponding to the minimum mapping size (line width of the drawing). φ_{xy} can be easily converted into the correspondent σ_{xy} by multiplying it per the nominal map scale.

For instance, the National Map Accuracy Standards for Horizontal Accuracy (Budget UB of the, 1947) assumes: (1) for map scale >1:20,000 ➔ $\varphi_{xy} = 0.508$ mm; (2) for map scale smaller or equal to 1:20,000 ➔ $\varphi_{xy} = 0.847$ mm. In this work φ_{xy} was set to 0.200 mm (Gomarasca, 2004). The following examples can help the reader to better get the point: if $\varphi_{xy} = 0.2$ mm, for a 1:10,000 scale map $\sigma_{xy} = 2$ m; for a 1:25,000 scale map σ_{xy} is 5 m, etc.

2.3. Modelling area uncertainty

A great variety of scientific applications use measurements (e.g. sizes, distances) coming from GIS tools operating on geographical data. Area computation is one of the most common operations that such applications require and, therefore, an associated value of its precision, is desirable. As previously mentioned, authors retain that an approach based on the modelling of the theoretical expected uncertainty is the most achievable one. The statistical tool for giving such an estimate is the variance propagation law (VPL, eq. 2). It enables the modelling of the relationship between the variance of direct measures (coordinates of polygon vertices) contributing to area computation and the variance of the area itself.

$$\sigma_y^2 = \left(\frac{\partial y}{\partial x_1}\right)^2 \bullet \sigma_{x_1}^2 + \left(\frac{\partial y}{\partial x_2}\right)^2 \bullet \sigma_{x_2}^2 + \ldots + \left(\frac{\partial y}{\partial x_n}\right)^2 \bullet \sigma_{x_n}^2 + 2\sum_{i=1}^{n-1} \times \sum_{j=i+1}^{n} \left(\frac{\partial y}{\partial x_i}\right)\left(\frac{\partial y}{\partial x_j}\right) COV(i,j) \quad (2)$$

where $y = f(x_1, x_2, \ldots, x_n)$ is the dependent variable, x_i the *i-th* independent variable and $\sigma_{x_i}^2$ its variance (supposed known); $COV(i,j)$ is the covariance between the *i-th* and *j-th* independent variables.

Direct measures involved in area (A) computation are the coordinates of polygon vertices (i.e. x, y). Area uncertainty can be therefore computed with reference to eq. 1 and eq. 2. If no significant correlation is assumed to exist between the independent variables, area uncertainty (σ_A) can be computed according to eq. 3.

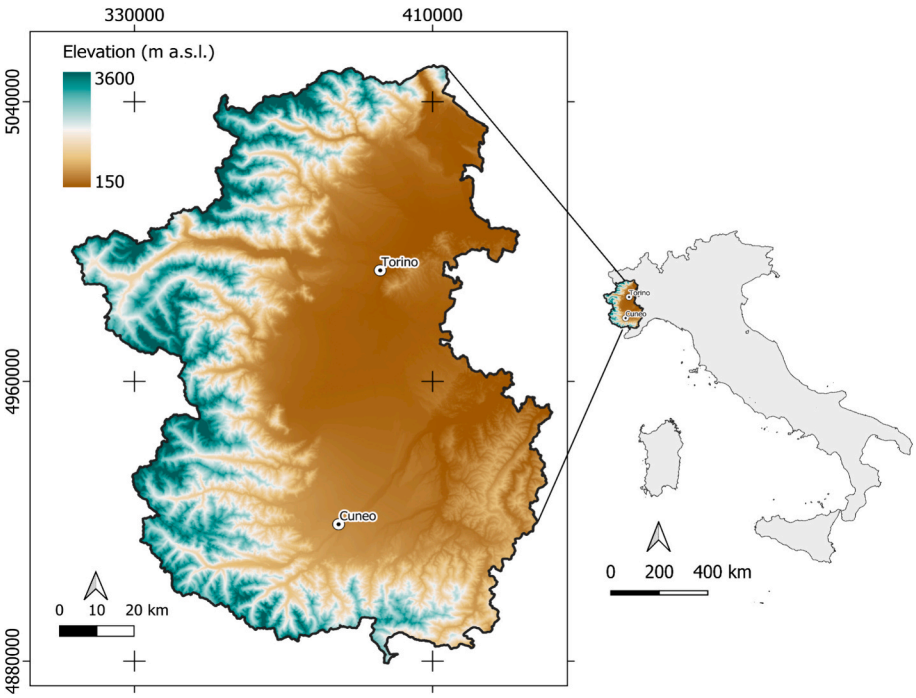


Fig. 2. AOI extension and terrain elevation. In grey Italian boundaries. Reference frame is World Geodetic System 1984 Universal Transverse Mercator 32 N (WGS84 / UTM32N).

$$\sigma_A = \sqrt{0.25 \bullet \sigma_{x,y}^2 \sum_{i=1}^{N} \left[(x_{i-1} - x_{i+1})^2 + (y_{i-1} - y_{i+1})^2 \right]} = \sqrt{0.25 \bullet \sigma_{x,y}^2 \sum_{i=1}^{N} l_{(i-1,i+1)}^2} \quad (3)$$

where N is the number of vertices; x_i, y_i are the *i-th* vertex coordinates; $l_{(i-1,i+1)}$ is the length of the vector linking the x and y coordinates of the vertices preceding and following the *i-th* one; $\sigma_{x,y}$ is the planimetric accuracy of the map, in this work set as $\sigma_{x,y} = \frac{\varphi_{xy} \bullet s}{\sqrt{2}}$ where s is the nominal map scale and $\varphi_{xy} = 0.2$ mm. In order to explore the theoretical sensitivity of σ_A to s and to the length of vectors linking polygon vertices, a self-developed routine was implemented in *R* vs *4.1.1.* All the possible combinations were tested by varying s in the range [10,1,000,000] and the quantity $\sum_{i=1}^{N} l_{(i-1,i+1)}^2$ in the range [10,1,000,000] m^2.

To make operational σ_A computation, the following procedure was applied in *QGIS* vs *3.16.11* (https://qgis.org) sequentially using built-in algorithms like field calculator or geometry conversions (Fig. 1): (i) initially, a unique identifier was assigned to each polygon; (ii) polygons were converted to lines and lines to segments (following polygon vertices); (iii) the length of segments was computed and recorded as new attribute in the layer table; (iv) the sum of squared lengths was computed according to the previously assigned polygon identifier and the result assigned to the correspondent polygon by table join; (v) given σ_{xy} (depending of the nominal scale of the map) eq. 3 was applied through the field calculator tool.

2.4. Suggesting operational implications

To make clear the operational implications of σ_A into the agro-forestry sector two paradigmatic case studies were reported assuming the same area of interest (AOI, Fig. 2). This includes two large Italian provinces, namely Torino (6827 km^2) and Cuneo (6905 km^2). They develop between 150 m and 3600 m (a.m.s.l.) and can be somehow assumed as representative of a typical Italian rural landscape.

2.4.1. Case study 1: Forest biomass mapping

Ordinarily forest biomass estimation is computed starting from ground surveys aimed at measuring (in few representative points) the unit biomass (i.e., m^3ha^{-1}). This has then to be referred to the entire forest area according to local forest types (Hunter et al., 2013; Somogyi et al., 2007). An error in forest area measure will therefore affect final biomass estimates since it results from unit biomass per forest area. For this analysis the Piemonte region forest map (FM) was adopted. FM is a polygon vector layer updated 2016 and representing the boundaries of forested areas in AOI. It maps >70 forest types (Camerano et al., 2017) corresponding to about 44,900 parcels (500,647 ha). Its nominal scale is 1:10000, therefore a $\sigma_{x,y}$ for this layer can be assumed equal to 2 m. The average value of forest biomass was estimated to be 175 m^3ha^{-1} (Gottero et al., 2007).

2.4.2. Case study 2: Common agricultural policy (CAP) payments implications

The CAP supports farming activities through economic grants with the aim of improving crop productivity, ensuring safe food production and contrasting climate change effects through a sustainable management of natural resources (Roederer-Rynning, 2010). In the CAP framework, since 2018 farmers are called to submit a Geo Spatial Aid Application (GSAA) to access the grants. GSAA include the interactive mapping of cultivated parcels by farmers that are merged to generate a map (AP, agricultural parcels) showing the actual agricultural context used along the CAP administrative procedures. AP correspond to a vector layer containing structured information about farmers, fields (land use, location, and size) and required grants (Sarvia et al., 2022a,b). In the most of cases the economic value of CAP contributions is related to the size of fields and, consequently to their area. To assess the effect of area uncertainty in CAP payments a total of 50,615 GSAAs were considered in AOI. Assuming AP nominal scale consistent with the one from Cadastre (i.e. 1:2000), the correspondent $\sigma_{x,y}$ can be set equal to 0.4 m. It is worth to outline that the average unit value of CAP payment (basic + greening payment) in AOI in 2022 was 300 € ha^{-1}.

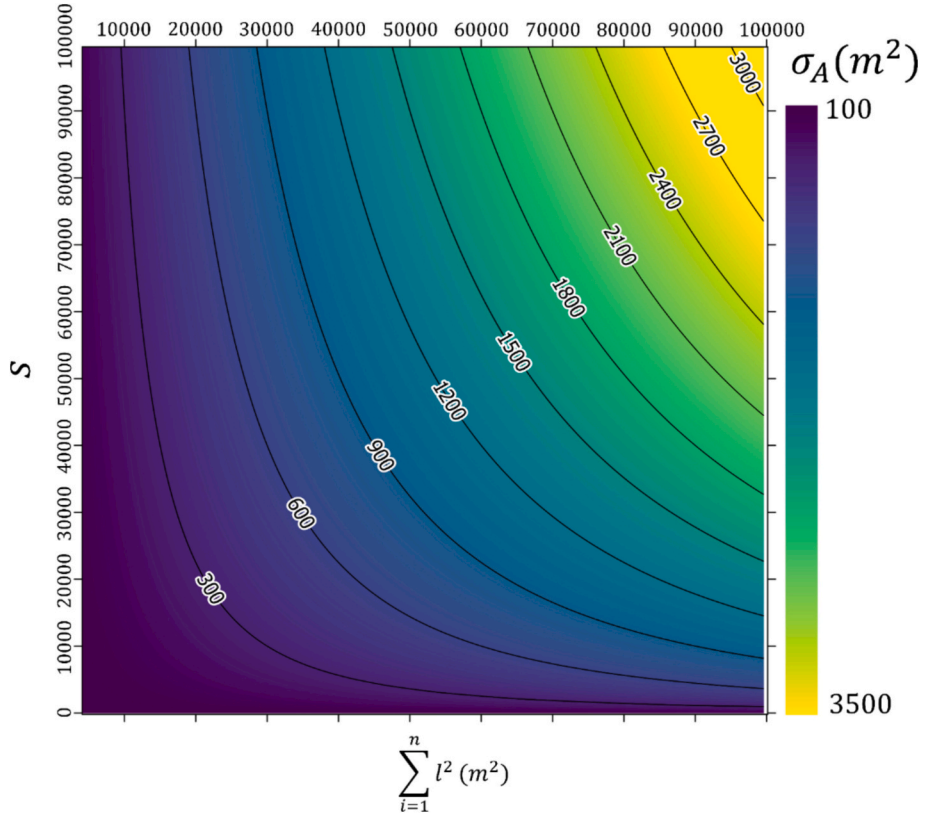


Fig. 3. Theoretical scenarios of σ_A obtained through VPL at different map scales and polygon sizes.

2.5. Area uncertainty vs polygon geometry

With reference to FM and AP, σ_A was computed in *QGIS* adopting the previously mentioned workflow (Fig. 1). The correspondent area relative uncertainty (u_{rel}) was also calculated as the ratio between σ_A and polygon area. The Kolmogorov-Smirnov (KS; Stephens, 1970) one-tailed test was used to compare the u_{rel} cumulative frequency distributions from the two dataset (CDF_{FM} and CDF_{AP}). This was intended for assessing how map scale (and therefore its precision) can affect the uncertainty of area computation. In particular, KS was built testing the following hypothesis: $CDF_{FM} > CDF_{AP}$.

To investigate if size and shape of polygon features somehow condition area accuracy some ordinary metrics from landscape ecology were used to qualify polygon (patch) shape. A wide range of metrics were defined (Baker and Cai, 1992; McGarigal and Marks, 1995; Riitters et al., 1995). For this work, five metrics were selected and tested against σ_A: (i) Area (A); (ii) Perimeter (P); (iii) number of vertices (V); (iv) Maximum Diameter (D_{max}); (v) Shape Index (*SI* [m^{-1}]). It is worth to remind that D_{max} represents the maximum distance between two vertices of a polygon (Lang and Blaschke, 2007); conversely, *SI* is a synthetic parameter describing shape complexity (Demetriou et al., 2013; Wentz, 1997). High *SI* values indicate complex, or elongated, shapes; low *SI* values indicate compact and isotropic shapes. SI is sensitive to size being inversely proportional to the polygon area.

The above mentioned metrics were computed for both FM and AP and related to σ_A through an ordinary multivariate regression Multiple linear regression analysis is frequently used to model the relationship linking a collection of predictors (x_j) to an outcome, or response variable (y). The "relative importance" of predictors can be explored through the so-called relative weights analysis (RWA). Researchers frequently look at the regression coefficients, or the zero-order correlations, to determine the relative importance of predictors (weights) (Gordon, 1968; Thompson and Borrello, 1985). Zero-order correlation quantify the contribution of the single predictor, with no regards about eventual other predictors. Differently, coefficients of regression define the amount of contribution given by a predictor to the response variable when it is combined with other predictors. If predictors are uncorrelated, their weights correspond to the R^2 from the related univariate regression. In this case, the sum of the single R_i^2 values from the *i-th* predictor is equal to the R^2 of the complete multivariate regression model. Unfortunately, predictors are frequently inter-correlated (multicollinearity) making regression coefficients inadequate (Budescu, 1993a). In this case, R_i^2 do not sum to R^2. Different RWA techniques have been therefore proposed to address this problem. The proposed solution is based on the possibility of measuring the amount of variance of the dependent variable that is explained by the single predictor participating to the multivariate model (Blackwell et al., 2000; LeBreton and Tonidandel, 2008; Tonidandel and LeBreton, 2010). This approach relies on the concept of "dispersion importance" (Anderson et al., 2006; Maliene et al., 2018), i.e. the proportion of the variance in y accounted for by x_j (predictor).

To take care of this potential situation, polygon metrics multicollinearity was verified according to the variance inflation factors analysis (Hsieh et al., 2003) and condition index (Senaviratna and Cooray, 2019). Given the existence of multicollinearity among predictors, their contribution to σ_A was estimated using 2 different RWA methods available in R, namely the "*olsrr*" (Hebbali and Hebbali, 2017) and "*relaimpo*" (Groemping and Matthias, 2018) packages. The adoption of two different methods was proposed to better support our deductions about the geometric features weights onto area uncertainty. The first method, proposed by Lindemann, Merenda and Gold (LMG) (Budescu, 1993b), is based on the computation of the squared semipartial correlation $r^2_{y(x_j|x_{1,\ldots,x_{j-1}})}$ (also called Type I, predictor variables added-in-order. The squared semipartial correlation for each succeeding predictor is then used to calculate the increment in the coefficient of determination at each stage. The Type I squared semipartial correlation value

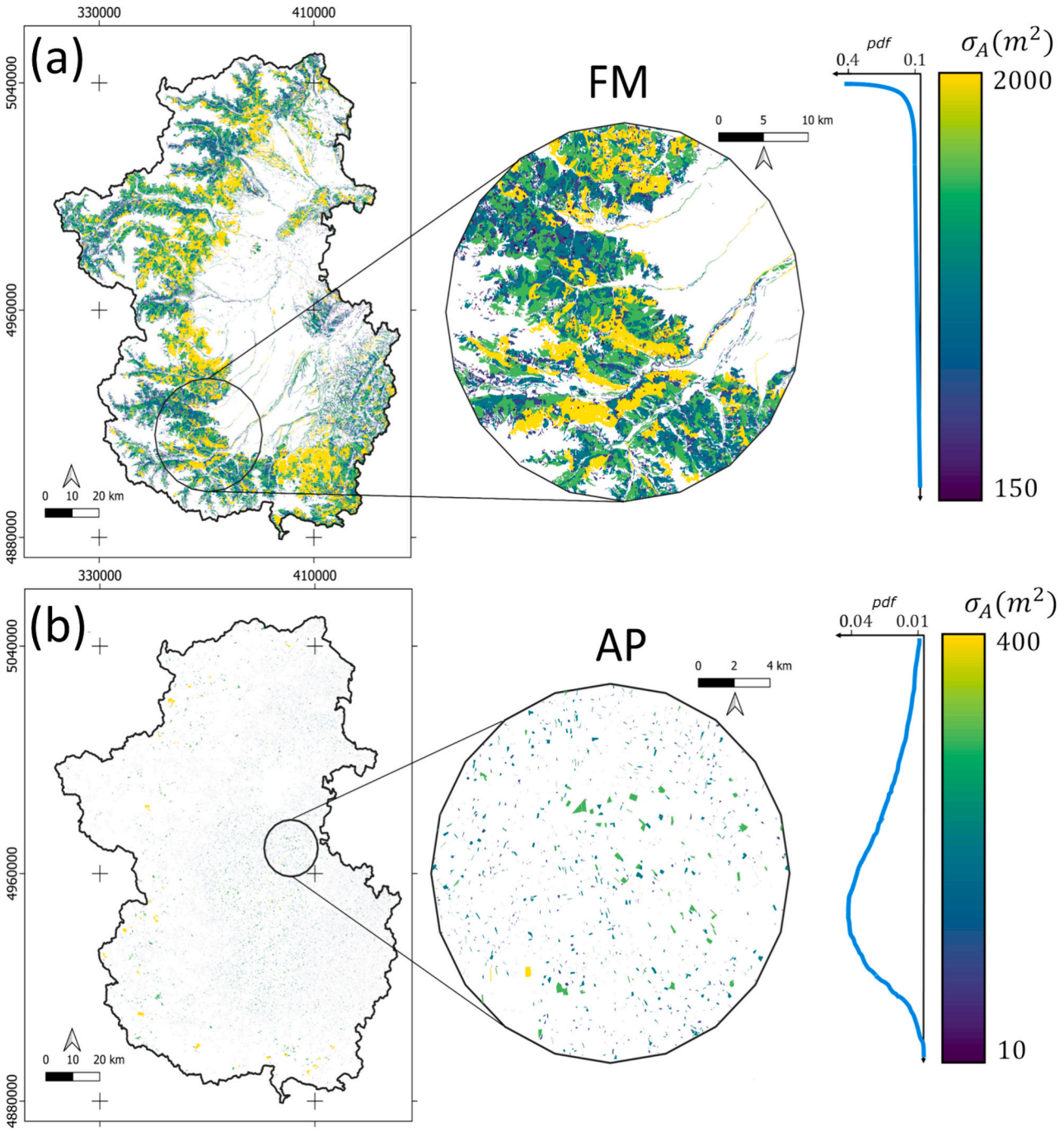


Fig. 4. (a) FM polygon area uncertainty map and related pdf; (b) AP polygon area uncertainty map and related pdf. Reference frame is WGS84/UTM 32 N.

for predictor x_j, however, relies on when it enters the model when the p predictors are mutually correlated. To take care about this problem, LMG computes the unweighted average of the $r^2_{y(x_j|x_{1,\ldots,}x_{j-1})}$ over all possible $p!$ orderings of how the p predictors can sequentially enter the model one-at-a-time (Chao et al., 2008). The second method is based on the covariance matrix decomposition and it is called the Johnson's Relative Weight (JRW) (Johnson, 2000). LMG and JWRI indices were compared to assess if weights estimates led to the same rank for predictors.

3. Results

3.1. Modelling area uncertainty

Uncertainty of polygon area computed through the Gauss's formula (eq. 1) was estimates suing VPL. In the first part of this work, a theoretical simulation was achieved to test sensitivity of σ_A to s and $\sum_{i=1}^{N} l^2_{(i-1,i+1)}$ values. s and $\sum_{i=1}^{N} l^2_{(i-1,i+1)}$ were iteratively changed in the range [10−1,000,000] and [10–1,000,000], respectively. Results are shown in Fig. 3 where it can be noted that σ_A increases more rapidly for smaller map scales when moving from small polygons to larger ones.

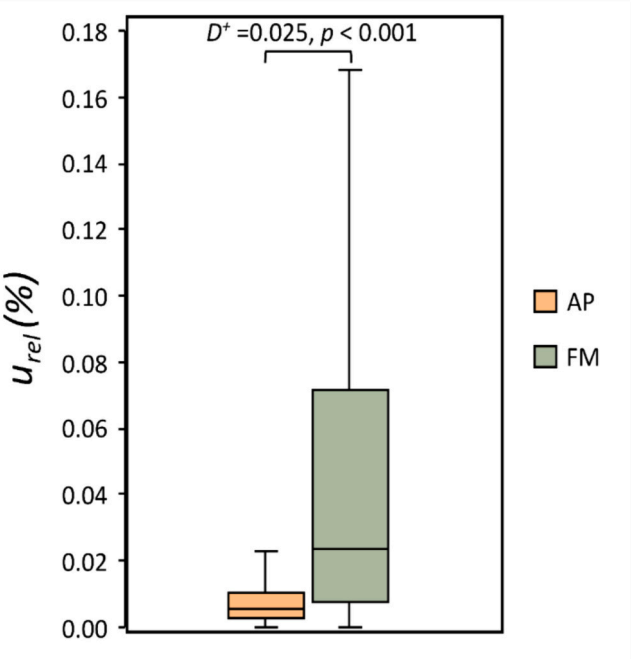


Fig. 5. Boxplots of FM and AM u_{rel}. KS test highlights how u_{rel} values distributions of FM is significantly greater than AP one.

3.2. Operational consequences

The workflow of Fig. 1 was implemented in *QGIS* environment and applied to two paradigmatic maps commonly used in the agro-foresty sector, namely FM and AP. Fig. 4a and b show spatial distribution of area uncertainty for FM and AP, respectively. Correspondent probability density function (*pdf*), namely pdf_{FM} and pdf_{AP}, are also reported. Comparing figures it can be easily noticed that: (i) agricultural (AP) and forestry (FM) landscape are characterised by polygons showing significantly different sizes. AM majorly contains small parcels, FM bigger ones; (ii) pdf_{FM} shows an exponential trend where more than the 40% of polygons present a σ_A value of 2000 m^2 while the remaining polygons are characterised by having a σ_A value between 150 and 2000 m^2. Conversely, considering pdf_{AP} (Fig. 4b), a mesokurtic distribution can be observed. Specifically, the σ_A is between 10 and 400 m^2, and it can be observed that most polygons are characterised by having a σ_A of around 100 m^2.

Fig. 5 shows the u_{rel} values distributions of both AP and FM. On average, FM's u_{rel} values are higher than AP's;specifically the average u_{rel} value results to be 0.01% and 0.025% for AP and FM respectively.

It is worth to note that in general small u_{rel} values exist (i.e., < 0.2%). KS test proved how u_{rel} values distributions of FM is significantly greater than AP one ($D^+ = 0.025$, $p < 0.001$) highlighting how, independently form polygon area, map scale can significantly affect u_{rel}. Nevertheless, a variability exists in u_{rel} values for both FM and AP (CV% was 7% and 15% respectively). This variability was further explored by assess how polygon geometric features can affect σ_A.

4. Exploring uncertainty vs polygon geometric features

A multivariate regression model was fitted for both AP and FM using ordinary least squares involving *A, P, V,* D_{max} and *SI* as predictors and σ_A as outcome variable. Concerning AP, the regression model resulted into a R^2 equal to 0.869 ($F = 59{,}770.5$, $p < 0.001$) while for FM a R^2 equal to 0.854 ($F = 59{,}209.1$, $p < 0.001$) was founded. Fig. 6 shows high correlation values among regression model predictors. Only SI resulted poorly correlated with both predictors and σ_A. Multicollinearity analysis performed using the condition index of FM and AP regression models resulted equal to 12.47 and 14.61 respectively highlighting a moderate multicollinearity and poor conditioned regression systems. Additionally, medium-high VIFs were founded (Fig. 6) highlighting a multicollinearity especially due to *P* and D_{max}. It is interesting how both FM and AP regression models shows very similar condition indices, VIFs and Pearson's correlation values, supporting the hypothesis that independently from the map scale and type of polygon, σ_A is affected by polygon geometric features. Nevertheless, not all geometric features equally influence σ_A. To explore how single predictors, affect polygon area error two different RWAs were performed.

In Fig. 7 are reported the relative weights of polygon geometric features onto σ_A. From this data, we can see that both LMG and JRW have generated the same predictor's importance ranking. This is notable since this ranking is independent from map scale and map type. In fact, Fig. 7a and b report similar weights. In particular, the most affecting geometric feature is the D_{max} accounting for the 50% of variance of σ_A in the FM context while it results to be around 40% for the agricultural one. The second one is the *P*, accounting for the 30% in both FM and AM context. These two geometric features summed can describe >80% and 70% of σ_A in the FM and AM context respectively. Otherwise, the smallest weight is the *SI*, accounting for <2%. *A* and *V* have very similar weight (between 10 and 15%).

5. Discussions

From the error model reported in eq. 3, we can note that polygon area uncertainty is directly proportional to the sum of squared segments lengths of polygon ($\sum_{i=1}^{N} l^2_{(i-1,i+1)}$) and one-quarter of the square positional error (σ^2_{xy}). This implies that polygons having equal size (area and perimeter) can show different area error according to the following interpretative key: less segments define the polygon higher is the area uncertainty. This dependency suggests how polygon digitalization process (photointerpretation/delineation) or image segmentation play a key role on σ_A and can greatly improve area-related estimates accuracy. Moreover, to explore the theoretical sensitivity of σ_A to both *s* and $\sum_{i=1}^{N} l^2_{(i-1,i+1)}$, all possible combinations were computed and σ_A scenarios reported in Fig. 4. It is interesting to note how polygons having similar $\sum_{i=1}^{N} l^2_{(i-1,i+1)}$ show a negative exponential behavior according to map scale. In fact, higher *s*-value (i.e., low detailed maps) higher the uncertainty. Unexpectedly, this behavior changed significantly with polygon having small $\sum_{i=1}^{N} l^2_{(i-1,i+1)}$. In fact, for small $\sum_{i=1}^{N} l^2_{(i-1,i+1)}$ values, σ_A is poorly affected by scale. For example, a polygon having $\sum_{i=1}^{N} l^2_{(i-1,i+1)} =$ 1000 m^2 shows σ_A equal to 300 m^2 for all scale values. While a polygon having $\sum_{i=1}^{N} l^2_{(i-1,i+1)} = 90{,}000$ m^2 shows σ_A values range of one-order of magnitude according to *s*-values (e.g., from 300 m^2 for large scale, up-to 3000 m^2 for small-scale). This suggest that in general area estimates from large polygons with few vertices should be carefully considered, especially if area-related estimates are adopted into operative workflows (i.e., for cadastral/tax purposes). RWA proved how σ_A is mainly affected by D_{max} and *P* independently from the adopted map. Higher values of these geometric features higher the area error. This is also proved by high positive correlation coefficients with σ_A (Fig. 6). This result may be explained by the fact that D_{max} and *P* are metrics perfectly related to $\sum_{i=1}^{N} l^2_{(i-1,i+1)}$; as supported also by theoretical scenarios, higher the lengths of the polygon segments higher the error. Interestingly, polygons having same D_{max} and *P* but more vertices can generate better area estimates. This outcome indicates that a more detailed polygon generated by a dense polygon drawing (many vertices) or by an automatic image segmentation procedure can lead to a better area result. Unexpectedly, no significant relationship exists between SI and σ_A. Ordinarily, SI describes how far a polygon is from an isotropic (circle) geometry, this implies that anisotropic polygons do not necessary result into a worst area estimate.

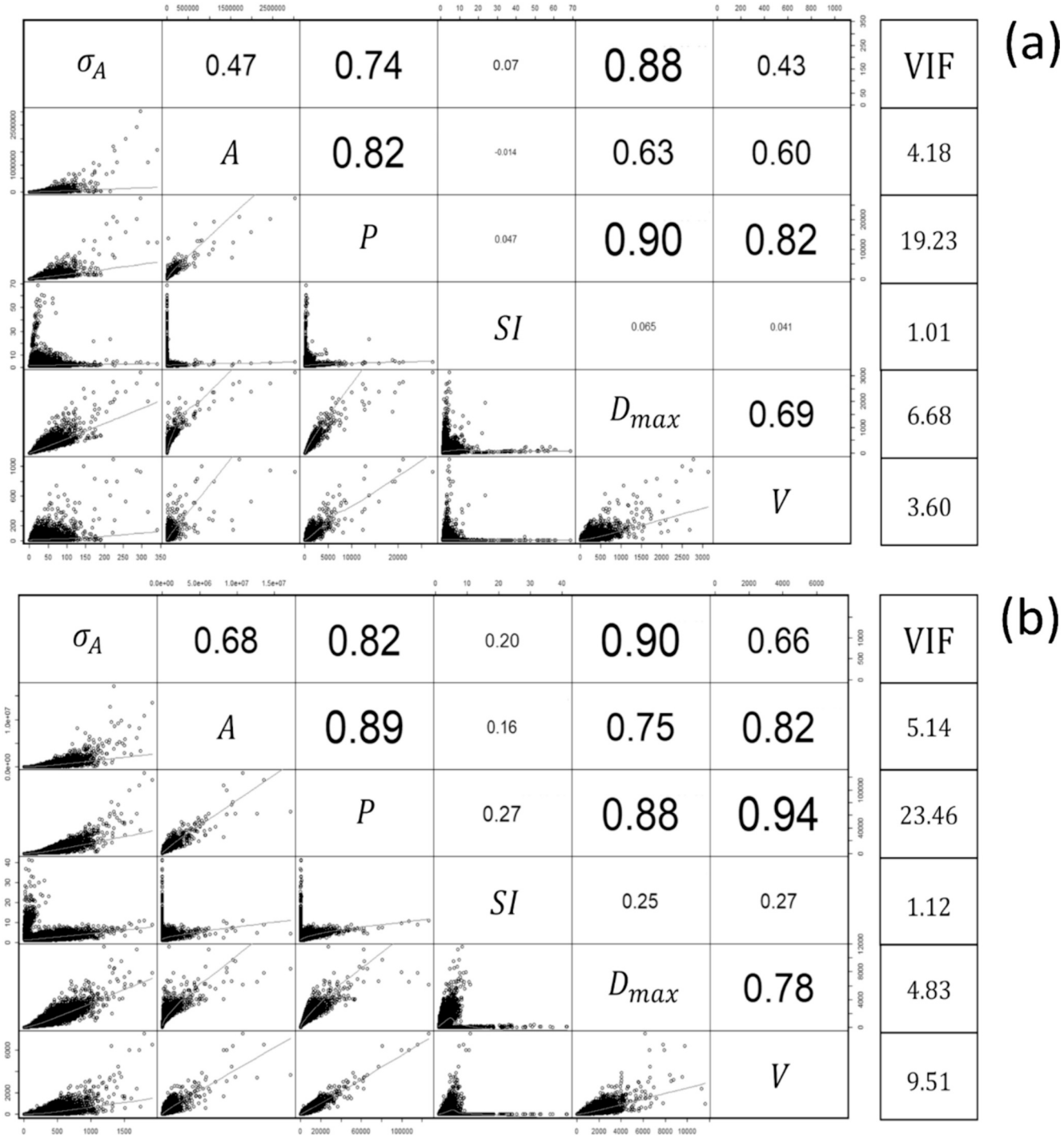


Fig. 6. Correlogram and VIF analysis of outcome (σ_A) and predictors variables involved in the multivariate regression models. (a) FM correlogram and VIFs; (b) AP correlogram and VIFs.

To give an idea how this outcome can affect deductions based on polygon vector data in the agro-forestry sector, in this work AP and FM were used. As mentioned in the previous sections, the agricultural sector and the correspondent farmers' activities are supported at European level through the CAP. Contributions value can be diverse and depend on the practices adopted by farmers. The amount of the contribution is not only influenced by the type of activity carried out, but also by the area on which the activity is applied. Consequently, not being aware of the uncertainty of a plot may result in a higher or lower payment than is really due. In this regard, given that AP u_{rel} is 0.01% (median value of Fig. 5) and that 50,615 fields with a total area of 21,103.8 ha were considered, the AOI σ_A of AP turns out to be 2.11 ha. Consequently, knowing that within AOI average CAP payments made in 2022 amounted to about 300 € ha^{-1}, and considering the 2.11 ha uncertainty derived from the calculation of the area of the plots, a value of 633.11 € may have been over- or underestimated. Naturally, this value is not very high, but the conditions change when all CAP applications made at European level are considered. For example, a total of 157 million hectares (belonging to approximately 9.1 million farms across Europe) were included in the CAP application for European subsidies in 2020 (www.ec.europa.eu). Consequently, by applying the results obtained within the study area to the European territory, keeping the value of

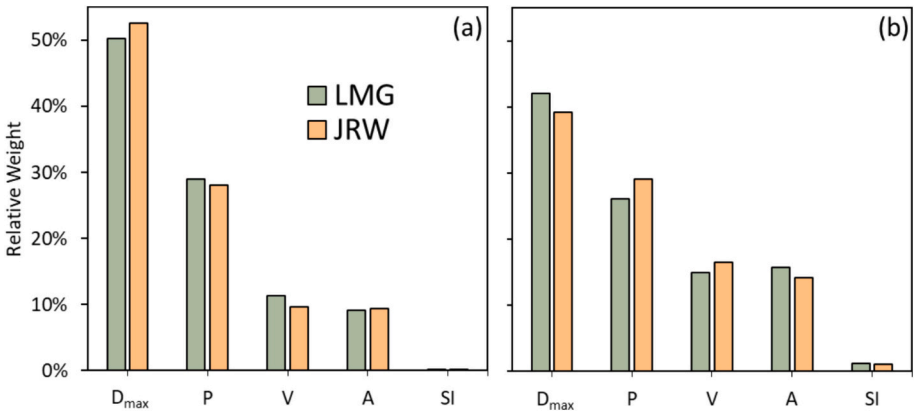


Fig. 7. RWA results according to LMG and JRW methods. (a) FM regression relative weights; (b) AP regression relative weights.

u_{rel} equal to 0.01%, the σ_A of AP at the European level turns out to be 15,700 ha. Therefore, applying a CAP contribution worth 300 € ha^{-1}, it can be assumed that a total of 4.7 million euro could be over- or underestimated.

Similar small AP u_{rel} values were found by Bogaert et al. (2005). They present a theoretical framework for assessing errors in area measurements of planar polygonal surfaces, suitable for both correlated and independent measurements, while working with CAP applications provided by farmers. Specifically, they found an area error ranging from 1% to 5% for typical EU field sizes alerting about the area uncertainty implications.

Considering the forest context, σ_A turns out to be also very important for biomass estimation, such parameter, for example, is widely used to derive the carbon sequestration estimation (Shi and Liu, 2017). Consequently, if the forest area uncertainty is not considered could compromise the biomass estimation and the corresponding carbon sequestration estimates. Moreover, as we have seen from the previous paragraphs, σ_A is strongly affected by the maximum diameter and perimeter, parameters that are often high in this context. In this regard, since u_{rel} is on average 0.025% and in the study area we have considered 44,900 fields, corresponding to a total area of 500,600 ha, the AOI σ_A of FP turns out to be about 125 ha. At this point, knowing that in AOI the average value of forest biomass is 175 m^3ha^{-1}, the biomass value that may have been over- or underestimated result to be equal to 21,900 m^3.

These small u_{rel} values are in line with those reported by previous studies on forest area uncertainty. In particular, Næset (1999) conducted research on a 717 ha forest in Åmot, southeast Norway, analyzing how errors by photointerpreters in defining forest stand boundaries affect area estimates, land use categorization, and total timber volume. Their findings indicate an average area underestimation of 2%. Moreover, they highlighted how these small area error propagates into a wood volume estimate error.

Finally, this work proposes and assess a method to compute the area uncertainty derived by polygon vector layer (map) obtained, possibly, by official providers whose generation is unknown. Rarely, metadata reports the survey technique (GNSS, LiDAR, photogrammetric processing, or editing from orthoimages) which is at the basis of map production. In this framework, a limitation of proposed method concerns the correlation of the measurement errors in the polygon vertexes as reported by some authors (De Bruin, 2008; Heuvelink et al., 2007). It is worth to additionally stress that this type of error strictly depends on survey geometric design making impossible to be properly modelled. Consequently, no hypothesis about spatial error correlation can be done. The only reasonable assumption we can do is to access map metadata and look for an explicit definition of the map horizontal accuracy or, if not present, to deduce it from the nominal scale of the provided data. Given these limitations, the proposed method assumes the uncertainty of vertexes positioning correspondent to the one defined in metadata or deducible by the nominal scale of the processed map. This is therefore propagated along the area formula, thus enabling the generalization of the method for all types of spatial data independently from the survey technique.

6. Conclusions

In this work a statistical based (i.e., VPL) approach to estimate the error involved in area calculations on polygon vector maps was proposed. Subsequently, a workflow was proposed and implemented in *QGIS* environment allowing to map σ_A and explore its spatial variability. Moreover, the relationship between area error and polygon geometric features was analyzed by using a multivariate regression and RWA. Results proved that two main factors condition polygon area calculation: the geometric accuracy of the image (or map) and the geometry of the polygon. It was demonstrated that polygons having longer segments generate low accurate area estimates that the same polygons with the same perimeter or area but having many segments. RWA highlighted how polygons having higher D_{max} and P values have higher σ_A. Finally, to highlight the operative implications while working with area error in agro-forestry sector, two case studies involving different maps (in terms of scale and polygon features) were explored. Results show that u_{rel} median values of about 0.01% and 0.02% were founded in AP and FM respectively. These errors can potentially alter the deductions based on polygon area measure with great operative consequences. Finally, the authors believe that an area uncertainty evaluation performed with proposed workflow and implemented in *QGIS* can provide meaningful information for both researchers and professionals in a wide range of applications, such as decision making, spatial data quality assessment and a proper quantification of area error can improve reliability of many results involving spatial data.

CRediT authorship contribution statement

Samuele De Petris: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Filippo Sarvia:** Writing – review & editing, Writing – original draft, Validation, Formal analysis. **Enrico Borgogno-Mondino:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors report there are no competing interests to declare.

Data availability

Data will be made available on request.

References

Alvarez-Mendoza, C.I., Teodoro, A., Ramirez-Cando, L., 2019. Improving NDVI by removing cirrus clouds with optical remote sensing data from Landsat-8–a case study in Quito, Ecuador. Remote Sens. Appl. Soc. Environ. 13, 257–274.

Amici, V., Maccherini, S., Santi, E., Torri, D., Vergari, F., Del Monte, M., 2017. Long-term patterns of change in a vanishing cultural landscape: a GIS-based assessment. Eco. Inform. 37, 38–51.

Anderson, M.J., Ellingsen, K.E., McArdle, B.H., 2006. Multivariate dispersion as a measure of beta diversity. Ecol. Lett. 9 (6), 683–693.

Baker, W.L., Cai, Y., 1992. The r. le programs for multiscale analysis of landscape structure using the GRASS geographical information system. Landsc. Ecol. 7 (4), 291–302.

Blackwell, B.G., Brown, M.L., Willis, D.W., 2000. Relative weight (Wr) status and current use in fisheries assessment and management. Rev. Fish. Sci. 8 (1), 1–44.

Bogaert, P., Delincé, J., Kay, S., 2005. Assessing the error of polygonal area measurements: a general formulation with applications to agriculture. Meas. Sci. Technol. 16 (5), 1170.

Bondesson, L., Ståhl, G., Holm, S., 1998. Standard errors of area estimates obtained by traversing and GPS. For. Sci. 44 (3), 405–413.

Budescu, D.V., 1993a. Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. Psychol. Bull. 114 (3), 542.

Budescu, D.V., 1993b. Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. Psychol. Bull. 114 (3), 542.

Budget UB of the, 1947. United States National Map Accuracy Standards.

Camerano, P., Terzuolo, P.G., Guiot, E., Giannetti, F., 2017. La Carta Forestale del Piemonte – Aggiornamento 2016. IPLA S.p.A. – Regione Piemonte.

Caspary, W., Scheuring, R., 1992. Error-band as measurers of geographic accuracy. In: Proceedings of EGIS'92.:226–233.

Chao, Y.-C.E., Zhao, Y., Kupper, L.L., Nylander-French, L.A., 2008. Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. J. Occup. Environ. Hyg. 5 (8), 519–529.

Chow, E., Kar, B., 2017. Error and accuracy assessment for fused data: remote sensing and GIS. In: Integrating scale in remote sensing and GIS, p. 125.

Chrisman, N.R., Yandell, B.S., 1988. Effects of point error on area calculations: a statistical model. Survey. Map. 48 (4), 241–246.

Chun, L., Xiaohua, T., 2005. Relationship of uncertainty between polygon segment and line segment for spatial data in GIS. Geo-spat. Inf. Sci. 8 (3), 183–188.

Crosetto, M., Tarantola, S., Saltelli, A., 2000. Sensitivity and uncertainty analysis in spatial modelling based on GIS. Agric. Ecosyst. Environ. 81 (1), 71–79.

De Bruin, S., 2008. Modelling positional uncertainty of line features by accounting for stochastic deviations from straight line segments. Trans. GIS 12, 165–177.

Demetriou, D., Stillwell, J., See, L., 2013. A GIS-based shape index for land parcels. In: First International Conference on Remote Sensing and Geoinformation of the Environment (rSCy2013), Vol. 8795. SPIE, place unknown, pp. 421–430.

Duarte, L., Teodoro, A.C., Gonçalves, H., 2014. Deriving phenological metrics from NDVI through an open source tool developed in QGIS. In: Earth Resources and Environmental Remote Sensing/GIS Applications V. SPIE, pp. 238–246.

Dutton, G., 1992. Handling positional uncertainty in spatial databases. In: Proceedings 5th International Symposium on Spatial Data Handling. [place unknown], pp. 460–469.

Gahegan, M., Ehlers, M., 2000. A framework for the modelling of uncertainty between remote sensing and geographic information systems. ISPRS J. Photogramm. Remote Sens. 55, 176–188.

Gomarasca, M., 2004. Elementi di geomatica. Associazione Italiana di Telerilevamento (AIT) Edizioni, Milano, p. 618.

Goodchild, M., Zhang, J., Kyriakidis, P., 2009. Discriminant models of uncertainty in nominal fields. Trans. GIS 13 (1), 7–23. https://doi.org/10.1111/j.1467-9671.2009.01141.x.

Gordon, R.A., 1968. Issues in multiple regression. Am. J. Sociol. 73 (5), 592–616.

Gottero, F., Ebone, A., Terzuolo, P., Camerano, P., 2007. I boschi del Piemonte, conoscenze e indirizzi gestionali. blu edizioni, Torino, IT.

Groemping, U., Matthias, L., 2018. Package 'relaimpo.' Relative importance of regressors in linear models (R package version).

Gross, C.-P., Adler, P., 1996. Reliability of area mapping by delineation in aerial photographs. In: United States Department Of Agriculture Forest Service General Technical Report Rm, pp. 267–271.

Hebbali, A., Hebbali, M.A., 2017. Package 'olsrr.' Version 05. 3.

Heuvelink, G.B., Burrough, P.A., Stein, A., 2006. Developments in analysis of spatial uncertainty since 1989. Class. IJGIS. 20, 91–95.

Heuvelink, G.B., Brown, J.D., van Loon, E.E., 2007. A probabilistic framework for representing and simulating uncertain environmental variables. Int. J. Geogr. Inf. Sci. 21 (5), 497–513.

Hsieh, F.Y., Lavori, P.W., Cohen, H.J., Feussner, J.R., 2003. An overview of variance inflation factors for sample-size calculation. Eval. Health Prof. 26 (3), 239–257.

Hunter, G.J., Qiu, J., Goodchild, M.F., 2000. Application of a new model of vector data uncertainty. In: Spatial Accuracy Assessment [Internet]. CRC Press, place unknown, pp. 203–208 [accessed 2024 Feb 22];. https://api.taylorfrancis.com/content/chapters/edit/download?identifierName=doi&identifierValue=10.1201/9781482279573-25&type=chapterpdf.

Hunter, M.O., Keller, M., Victoria, D., Morton, D.C., 2013. Tree height and tropical forest biomass estimation. Biogeosciences 10 (12), 8385–8399.

Johnson, J.W., 2000. A heuristic method for estimating the relative weight of predictor variables in multiple regression. Multivar. Behav. Res. 35 (1), 1–19.

Judge, P.A., Allmendinger, R.W., 2011. Assessing uncertainties in balanced cross sections. J. Struct. Geol. 33 (4), 458–467.

Kiiveri, H.T., 1997. Assessing, representing and transmitting positional uncertainty in maps. Int. J. Geogr. Inf. Sci. 11 (1), 33–52.

Ku, H.H., 1966. Notes on the use of propagation of error formulas. J. Res. Natl. Bur. Stand. 70 (4).

Lang, S., Blaschke, T., 2007. Landschaftsanalyse mit GIS [place unknown]:. Ulmer Stuttgart.

LeBreton, J.M., Tonidandel, S., 2008. Multivariate relative importance: extending relative weight analysis to multivariate criterion spaces. J. Appl. Psychol. 93 (2), 329.

Leung, Y., Ma, J.-H., Goodchild, M.F., 2004. A general framework for error analysis in measurement-based GIS part 4: error analysis in length and area measurements. J. Geogr. Syst. 6 (4), 403–428.

Maliene, V., Dixon-Gough, R., Malys, N., 2018. Dispersion of relative importance values contributes to the ranking uncertainty: sensitivity analysis of multiple criteria decision-making methods. Appl. Soft Comput. 67, 286–298.

McGarigal, K., Marks, B.J., 1995. FRAGSTAT. In: Spatial analysis program for quantifying landscape structure. USDA Forest Service General Technical Report PNW-GTR-351.

Mikhail, E.M., Ackermann, F.E., 1976. Observations and Least Squares. IEP, New York.

Næset, E., 1999. Effects of delineation errors in forest stand boundaries on estimated area and timber volumes. Scand. J. For. Res. 14 (6), 558–566.

Neteler, M., Bowman, M., Land, H., Metz, M., 2012. GRASS GIS: a multi-purpose open source GIS. Environ. Model. Softw. 31, 124–130. Links.

Nowak, M.M., Skowroński, J., Slupecka, K., Nowosad, J., 2023. Introducing tree belt designer-A QGIS plugin for designing agroforestry systems in terms of potential insolation. Eco. Inform. 75, 102012.

Perkal, J., 1956. On epsilon length. Bull. l'académie Polonaise Sci. 4 (3), 399–403.

Rae, C., Rothley, K., Dragicevic, S., 2007. Implications of error and uncertainty for an environmental planning scenario: a sensitivity analysis of GIS-based variables in a reserve design exercise. Landsc. Urban Plan. 79 (3–4), 210–217.

Riitters, K.H., O'neill, R.V., Hunsaker, C.T., Wickham, J.D., Yankee, D.H., Timmins, S.P., Jones, K.B., Jackson, B.L., 1995. A factor analysis of landscape pattern and structure metrics. Landsc. Ecol. 10 (1), 23–39.

Roederer-Rynning, C., 2010. The common agricultural policy. In: Policy-Making in the European Union. Oxford University Press, place unknown, pp. 181–205.

Sarvia, S., De Petris, S., Borgogno-Mondino, E., 2022a. Detection and counting of meadow cuts by copernicus sentinel-2 imagery in the framework of the common agricultural policy (CAP). Eur. J. Remote Sens. 1–15.

Sarvia, F, De Petris, S, Ghilardi, F, Xausa, E, Cantamessa, G, Borgogno-Mondino, E, 2022b. The Importance of Agronomic Knowledge for Crop Detection by Sentinel-2 in the CAP Controls Framework: A Possible Rule-Based Classification Approach. Agronomy 12 (5), 1228. https://doi.org/10.3390/agronomy12051228.

Sarvia, F., De Petris, S., Borgogno-Mondino, E., 2023. Mapping melliferous potential in productive honey areas through spatial tools: towards a rationalization of beekeeping. Eco. Inform. 78, 102362.

Senaviratna, N., Cooray, T., 2019. Diagnosing multicollinearity of logistic regression model. Asian J. Probab. Stat. 5 (2), 1–9.

Shi, W., 1998. A generic statistical approach for modelling error of geometric features in GIS. Int. J. Geogr. Inf. Sci. 12 (2), 131–143.

Shi, W., 2009. Principles of Modeling Uncertainties in Spatial Data and Spatial Analyses. CRC press, place unknown.

Shi, L., Liu, S., 2017. Methods of estimating forest biomass: a review. Biomass Volume Estimat. Valoriz. Energy. 10, 65733.

Shi, W.Z., Ehlers, M., Tempfli, K., 1999. Analytical modelling of positional and thematic uncertainties in the integration of remote sensing and geographical information systems. Trans. GIS 3 (2), 119–136.

Somogyi, Z., Cienciala, E., Mäkipää, R., Muukkonen, P., Lehtonen, A., Weiss, P., 2007. Indirect methods of large-scale forest biomass estimation. Eur. J. For. Res. 126 (2), 197–207.

Stephens, M.A., 1970. Use of the Kolmogorov–Smirnov, Cramer–Von Mises and related statistics without extensive tables. J. R. Stat. Soc. B. Methodol. 32 (1), 115–122.

Suwanlee, S.R., Pinasu, D., Som-ard, J., Borgogno-Mondino, E., Sarvia, F., 2024. Estimating sugarcane aboveground biomass and carbon stock using the combined time series of sentinel data with machine learning algorithms. Remote Sens. 2024 (16), 750. https://doi.org/10.3390/rs16050750.

Temme, A., Heuvelink, G.B.M., Schoorl, J.M., Claessens, L., 2009. Geostatistical simulation and error propagation in geomorphometry. Dev. Soil Sci. 33, 121–140.

Thompson, B., Borrello, G.M., 1985. The importance of structure coefficients in regression research. Educ. Psychol. Meas. 45 (2), 203–209.

Tonidandel, S., LeBreton, J.M., 2010. Determining the relative importance of predictors in logistic regression: an extension of relative weight analysis. Organ. Res. Methods 13 (4), 767–781.

van Oort, P.A.J., Stein, A., Bregt, A.K., De Bruin, S., Kuipers, J., 2005. A variance and covariance equation for area estimates with a geographic information system. For. Sci. 51 (4), 347–356.

van Oort, Stein, A., Bregt, K., de Bruin, S., Kuipers, J., 2005. A Variance and Covariance Equation for Area Estimates with a Geographic Information System. Forest Science 51 (4), 347–356. https://doi.org/10.1093/forestscience/51.4.347.

Wentz, E.A., 1997. Shape analysis in GIS. In: Proceedings of Auto-Carto, Vol. 13, pp. 7–10 [place unknown].
Zhou, L., Stein, A., 2013. Application of random sets to model uncertainty of road polygons extracted from airborne laser points. Comput. Environ. Urban. Syst. 41, 289–298.
Zubaer, K.H., Alam, Q.M., Toha, T.R., Salim, S.I., Al Islam, A.A., 2020. Towards simulating non-lane based heterogeneous road traffic of less developed countries using authoritative polygonal GIS map. Simul. Model. Pract. Theory 105, 102156.