International Journal of Remote Sensing

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tres20

A predictive modelling technique for human population distribution and abundance estimation using remote-sensing and geospatial data in a rural mountainous area in Kenya

M. Siljander [a], B.J.F. Clark [a] & P.K.E. Pellikka [a]

[a] Department of Geosciences and Geography, University of Helsinki, P.O. Box 64, 00014, Helsinki, Finland

Available online: 10 Aug 2011

To cite this article: M. Siljander, B.J.F. Clark & P.K.E. Pellikka (2011): A predictive modelling technique for human population distribution and abundance estimation using remote-sensing and geospatial data in a rural mountainous area in Kenya, International Journal of Remote Sensing, DOI:10.1080/01431161.2010.499383

To link to this article: http://dx.doi.org/10.1080/01431161.2010.499383

International Journal of Remote Sensing
iFirst, 2011, 1–27

A predictive modelling technique for human population distribution and abundance estimation using remote-sensing and geospatial data in a rural mountainous area in Kenya*

M. SILJANDER**, B. J. F. CLARK and P. K. E. PELLIKKA
Department of Geosciences and Geography, University of Helsinki,
P.O. Box 64, 00014 Helsinki, Finland

(*Received 31 December 2008; in final form 27 November 2009*)

This study presents a predictive modelling technique to map population distribution and abundance for rural areas in Africa. Prediction models were created using a generalized regression analysis and spatial prediction (GRASP) method that uses the generalized additive model (GAM) regression technique. Dwelling unit presence–absence was mapped from airborne images covering 98 km^2 (30% of the study area) and used as a response variable. Remote-sensing-based (reflectance, texture and land cover) and geospatial (topography, climate and distance) data were used as predictors. For the rest of the study area (228 km^2; 70%), GAM models were extrapolated, and prediction maps constructed. Model performance was measured as explanatory power (adj.D^2, adjusted deviance change), predictive power (area under the receiver operator curve, AUC) and kappa value (κ). GAM models explained 19–31% of the variation in dwelling-unit occurrence and 28–47% of the variation in human population abundance. The predictive power for population distribution GAM models was good (AUC of 0.80–0.86). This study shows that for the prediction of dwelling-unit distribution and for human population abundance, the best modelling performance was achieved using combined geospatial- and remote-sensing-based predictor variables. The best predictors for modelling the variability in human population distribution using combined predictors were angular second moment image-texture measurement, precipitation, mean elevation, surface reflectance for Satellite Pour l'Observation de la Terre (SPOT) red and near-infrared (NIR) bands, correlation image-texture measurement and distance to roads, respectively. The population-abundance modelling result was compared with two existing global population datasets: Gridded Population of the World version 3 (GPWv3) and LandScan 2005. The result showed that for regional and local-scale population-estimation probability, models created using remotely sensed and geospatial data were superior compared to GPWv3 or LandScan 2005 data products. Population models had high correlation with Kenyan population census data for 1999 in mountainous sub-locations and low correlation for sub-locations that also extended into the lowlands.

*This paper came from a workshop entitled '*Potentialities and Limitations in the Use of Remote Sensing for Detecting and Monitoring Environmental Change in the Horn of Africa*'. The workshop was organized by Somalia Water and Land Information Management (SWALIM) between 13th and 14th June 2007 at Holiday Inn in Nairobi, Kenya. SWALIM is a project of the UN-FAO in Somalia (www.faoswalim.org).

**Corresponding author. Email: mika.siljander@helsinki.fi

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online © 2011 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/01431161.2010.499383

1. Introduction

Africa is experiencing one of the highest population growth rates in the world. The sub-Saharan population is growing at a rate of 2.5% per year according to the United Nations Children's Fund (UNCEF), and in Kenya, both the growth rate of 2.8% and the total fertility rate of 5.0 are extremely high. Population has grown from 10.9 million in 1969 to 28.7 million in 1999 and, according to the United Nations (UN), population is estimated to grow to 46 million by 2015. The population growth has significant social, economic and environmental consequences, and there is an urgent need to deepen our understanding of population distribution and population-abundance patterns. Demographic information is usually provided in national or administrative units, and for local-scale population analysis, more accurate measurement units are needed. Furthermore, these sub-national reference units can be vastly different in size and shape. For spatial analysis, it is often preferable to record human population estimates using standardized units, such as regular analysis grids (Mubareka *et al.* 2008). However, most grid-based human population models (e.g. LandScan; Dobson *et al.* 2000) exist only at a coarse scale, thus generalizing and obscuring the internal variability of population data. The larger the size of the analysis grid unit, the more generalized the data are and the less suitable they are to be used at regional and especially local scales. Therefore, cost-efficient applications to create spatially explicit human population geospatial databases and distribution maps at finer scales are in great demand.

Aerial photography has been the traditional method to estimate and to map population distribution at local and regional scales. Porter (1956) used a rural dwelling-unit count in Liberia, and Hsu (1971) used the same method in the Atlanta and Boston metropolitan areas in the USA. A number of other studies have shown that a dwelling-unit count from aerial photography is the most accurate remote-sensing method for population estimation and distribution mapping (e.g. Lindgren 1971, Forester 1985, Lo 1986, 1995). Unfortunately, the conventional dwelling-unit count method is a time-consuming expensive process, and it requires abundant high-resolution aerial photographs to cover large areas (Lo 1989). For rural areas in developing countries such as Kenya, up-to-date high-resolution airborne imagery is typically lacking. Modern geospatial techniques and data, however, provide new possibilities for population distribution estimation in such areas without the need for high-resolution airborne remote-sensing data. Powerful remote-sensing and geographic information system (GIS) tools and statistical techniques can now be used for predictive modelling to test the hypothesis that human population distribution and abundance can be predicted using remotely sensed and geospatial-based predictors.

Predictive modelling techniques have been used successfully for over a decade in ecological applications, such as in species-distribution and abundance modelling (Guisan and Zimmermann 2000, Lehmann *et al.* 2002). This is mainly due to the fact that wildlife population distribution and abundance can be predicted based on habitat requirements of a given species, such as vegetation cover, distance to water, temperature, precipitation and incoming solar radiation. It is more challenging to predict human population distribution and abundance as humans have the capability to modify the environment as well as to transport elements of 'suitable habitat' from remote locations through trade and exchange. This could explain why predictive techniques have been used more rarely for human population distribution and abundance modelling. For existing human population models, mainly the linear least squares (LS) regression method has been deployed (see, e.g. Schnaiberg

et al. 2002, Gustafson *et al.* 2005, Li and Weng 2005). In these models, environmental predictors have been used to explain the variation in dwelling-unit or human population data. However, the linear LS regression method has implicit statistical assumptions; for example, the assumption of linearity, independence, homogeneity of variance, and normality (Zar 1999), which are often violated when environmental data with non-Gaussian and non-constant variance are used in regression analysis. To counter the regression analysis violations discussed above, we used generalized additive models (GAMs) for prediction. This regression method supports non-Gaussian error distributions and non-linear relationships between response and predictor variables (Hastie and Tibshirani 1990). To the best of our knowledge, GAMs have not previously been used for human population prediction modelling, and therefore we developed a hybrid modelling method using a 100 m analysis square size, advanced remote-sensing, geospatial and statistical data derived from airborne remote-sensing data, Satellite Pour l'Observation de la Terre (SPOT) satellite data and a geospatial database. The resolution is considered suitable for local-scale population studies and other GIS applications. We used the generalized regression analysis and spatial prediction (GRASP) modelling method (Lehmann *et al.* 2002) that exploits GAMs in a semi-automatic manner and has been used previously in species-distribution prediction analyses (e.g. Zaniewski *et al.* 2002, Maggini *et al.* 2006).

In previous predictive human population studies, several predictors have proven to be superior for determining population distribution. Land cover has been one of the main factors used for determining population abundance (Dobson *et al.* 2000). However, classified land cover, derived from moderate-resolution satellite imagery such as the Landsat Enhanced Thematic Mapper Plus (ETM+), has been criticized to be an inadequate predictor because it aggregates the true land cover (St-Louis *et al.* 2006, Bellis *et al.* 2008). An alternative method to thematic land-cover data is to use image-texture values as remote-sensing-based predictors; for example, those based on the grey-level co-occurrence matrix (GLCM) derived directly from satellite imagery (Haralick *et al.* 1973, Li and Weng 2005). Such measures can be used to quantify the variability of vegetation as a continuous variable in statistical modelling (Bellis *et al.* 2008). In addition to land cover and image-texture predictors, we used environmental predictors, derived from geospatial datasets and models, which are widely agreed to be the best predictors for human population distribution and abundance prediction (Dobson *et al.* 2000). These predictors are elevation, aspect, slope, precipitation, distance to roads and distance to water. Topographic wetness index (TWI) and solar radiation energy (irradiance) were also used as predictors. Each of the predictors was derived from the Taita Hills Environmental Monitoring System (THEMS) database. In the modelling, we used three types of predictor groups to reveal which type of group had the best performance in explaining the variation in dwelling-unit presence–absence and abundance data. The predictor groups were: remote-sensing-based, geospatial and combined predictors groups. A description of the predictors can be seen in table 1, and they are discussed in more detail in section 3.

We purposely did not attempt to include socio-economic factors in the models, except one surrogate predictor 'the distance calculated to the main roads', because of five reasons: firstly, socio-economic factors are not available for the study area in a reference unit that would be usable in this study; secondly, the reliability of the data is questionable; thirdly, socio-economic factors, except distance-based calculations, are also lacking from other remote-sensing- and GIS-based human population prediction studies; fourthly, we used remote-sensing and geospatial predictors that can

Table 1. Predictors from geospatial data and remote-sensing data used to generate the human population distribution and abundance models. Variables correspond to percentage or means obtained from averaging individual values from pixels to 100 m analysis square.

Predictor	Description
Predictors from geospatial data	
Elevation	Mean elevation from DEM (Survey of Kenya 1: 50 000 scale topographic map)
Aspect	Mode aspect in degrees in 100 m analysis square (derived from DEM)
Slope	Mean slope degree in 100 m analysis square (derived from DEM)
TWI	Mean topographical wetness index in 100 m analysis square (derived from DEM)
Irradiance	Mean annual irradiance (kW h m^{-2}) (scaled between 0 to 1) in 100 m analysis square (derived from DEM)
Precipitation	Mean annual precipitation (mm) using ANUSPLIN interpolation method
Distriver	Mean Euclidean distance to rivers (m) in 100 m analysis square.
Distroad	Mean Euclidean distance to roads (m) in 100 m analysis square.
Predictors from remote-sensing data	
Reflectance, red band	Mean spectral reflectance for SPOT satellite imagery red band
Reflectance, NIR band	Mean spectral reflectance for SPOT satellite imagery NIR band
Asm2	Mean angular second moment image texture for SPOT band 2 using 7×7 analysis window
Corr2	Mean correlation image texture for SPOT band 2 using 7×7 analysis window
Croplands	Percentage of crops land-cover class in 100 m analysis square using SPOT imagery LULC mapping
Thicket	Percentage of thicket land-cover class in 100 m analysis square using SPOT imagery LULC mapping
Woodland	Percentage of woodland land-cover class in 100 m analysis square using SPOT imagery LULC mapping
Plantation forest	Percentage of plantation forest land-cover class in 100 m analysis square using SPOT imagery LULC mapping

Note: NIR, near-infrared; DEM, digital elevation model; LULC, land-use land-cover.

be obtained and derived easily, thus making this type of modelling technique suitable for application to other areas; and fifthly, our aim was to examine the potential of remote-sensing and geospatial predictors for human distribution and abundance predictive mapping. Therefore, our models are based on environmental predictors derived from spatially explicit geospatial data using GIS and remote-sensing techniques.

To validate our models, we used a comparative analysis between predicted human abundance GAM models and two coarse-scale population datasets: the latest version of Gridded Population of the World (GPWv3) at 5 km resolution and LandScan 2005 (Dobson *et al.* 2000) at 1 km resolution. The GPW project is a continuum to the work of Tobler *et al*. (1995, 1997) to create a world population map with 5′ resolution based on sub-national census data and a smoothing algorithm. The latest version of the GPWv3 dataset is available in 2.5′ resolution (Balk and Yetman 2004) from the Center for International Earth Science Information Network (CIESIN) Internet site (CIESIN 2005). The LandScan project (Dobson *et al.* 2000) is another

global population dataset at 30″ resolution. The dataset is built using a model that distributes sub-national census data into grids by using various remote-sensing and geospatial data layers taking into account road proximity, slope, land cover and night-time lights, and is validated using high-resolution panchromatic imagery. The newest version, LandScan 2008, is available from the LandScan Internet site. A third effort to make a gridded population dataset (in 1° resolution for the year 2000) has been carried out under the UN Environment Programme/Global Resource Information Database (UNEP/GRID 2006). This model was not used in the present study because of the very coarse resolution and the time difference in creation of the database. We also compared all three grid-based population models, prediction models, GPWv3 and LandScan 2005 with data derived from the Kenyan population census in 1999 (Republic of Kenya 2001).

Human population is an important component in landscape ecological studies and geospatial modelling, but the currently available datasets have their limitations. The resolution of the global population datasets, 5 km resolution of GPWv3 and 1 km resolution of LandScan 2005, is too coarse for a very diverse and fragmented rural mountainous landscape study area. Furthermore, the Kenyan population census is undertaken with a 10 year interval and gives population for sub-locations as units, which are often coarse and inappropriate for use in local-scale studies. The last census in Kenya was in 1999 and, with population growth of 1.8% in the Taita Hills study area, the data are outdated for many applications. In addition, as the sub-locations are large, extending from densely inhabited hills to sparsely populated dry plains, the geographical units are inadequate to model the population distribution in a given landscape. To overcome these limitations in the current population datasets, we employed a predictive technique to create 100 m resolution spatially explicit human population distribution and abundance geospatial models for the Taita Hills. These fine-scale gridded human population distribution and abundance maps take into account internal variability of population data and are therefore more suitable to be used in local-scale environmental studies, such as biodiversity (Githiru and Lens 2004), land-use and land-cover change (Pellikka *et al.* 2009) and other studies in the Taita Hills.

The objectives of the study were:

1. to determine how well human population distribution and abundance can be predicted by predictors derived from remote-sensing and geospatial data and to determine which are the best predictors;
2. to explore the potential of remote sensing for enhancing the predictive power of human population distribution and abundance models;
3. to extrapolate the predictive human population distribution and abundance map for the whole of the Taita Hills; and
4. to compare the results with two existing population datasets and Kenya population census data of 1999.

2. Study area and human population in the Taita Hills

The Taita Hills are located in the Taita Taveta district of southeastern Kenya at 03° 25′ S, 38° 20′ E (figure 1). The elevation limit to separate hills from lowlands was set to 1100 m above sea level (a.s.l.), based on the gradient change from the plains to foothills, resulting in a study area of 326 km^2. Annual precipitation in the hills is 1200 mm, whilst in the plains, it is c. 600 mm based on records from the Kenya

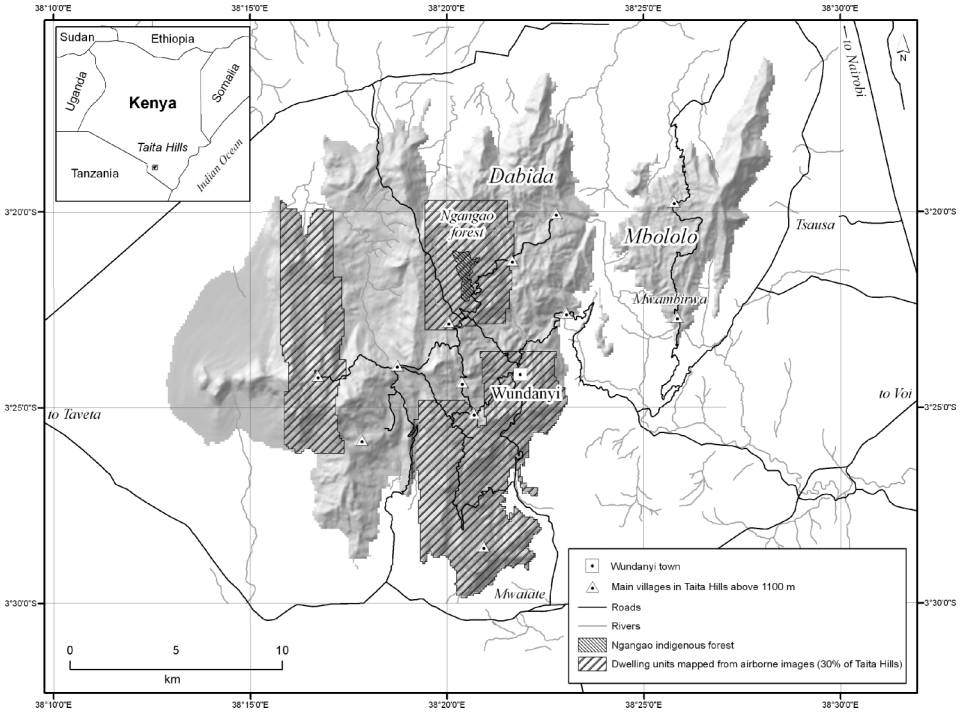

Figure 1. The Taita Hills in southeastern Kenya. The areas above 1100 m a.s.l. are presented as a shaded relief model, and the model building areas (30% of the Taita Hills), from where the dwelling units were digitized, are shown as hatched polygons.

Meteorological Institute. The southeastern and northeastern trade winds bring in moisture from the Indian Ocean and cause orographic rains on the eastern slopes; as a result, the western and northern slopes are in rain shadow. Population in the Taita Taveta district was 246 671 inhabitants (Republic of Kenya 2001) in 1999, and it is concentrated in the fertile Taita Hills, Sagala Hills and the town of Taveta close to Mt. Kilimanjaro and in trading centres, such as Voi and Mwatate. The population in the Taita Hills is concentrated in the best agricultural areas close to fertile river valleys in areas receiving rainfall more than 1000 mm annually, with an exception that the Mbololo massif is less densely populated compared to the Dabida massif.

The district capital Wundanyi is the only town in the hills with c. 4500 inhabitants in Wundanyi sub-location in 1999, whilst the rural area consists of villages of varying size. The population density in the best agricultural areas is between 400 and 500 inhabitants per km^2, but some sub-locations have a density of more than 900 persons according to the 1999 Kenya census. The sub-locations consisting of areas in the hills and the lowlands have a density between 100 and 200 people per km^2, whilst the rural lowlands have a density between 5 and 30 people. The population in Wundanyi, Mwatate, Tausa and Mwambirwa divisions was 135 000 based on the 1999 census, in an area of 3000 km^2, resulting in an average population density of 253 per km^2. The least populated areas are the western and northern plains and northwestern parts of the Taita Hills due to the rain-shadow effect. Most people in the hills live on small farms at elevations between 1300 and 1800 m a.s.l., where the annual long-term rainfall

varies between 360 and 1935 mm, and the mean daytime temperature is 20.5°C. The average size of a farm is 2 ha (Soini 2005, Ruotsalainen 2008).

3. Material and methods

We used geospatial data and the GRASP regression method for predictive modelling. We tested the hypothesis that remotely sensed and geospatial predictors can be used to predict human population distribution and abundance. For the model building, the response variable (dwelling-unit presence–absence) was derived from airborne imagery covering 30% of the Taita Hills. Geospatial GIS- and remote-sensing-based data layers were used as predictors. We randomly divided model building data ($n = 10\,488$, 100 m analysis squares) into model calibration 70% ($n = 7342$) and model evaluation 30% ($n = 3146$) datasets (n = sample size). Prediction models were extrapolated to cover the whole Taita Hills area ($n = 34\,143$) and the human population-abundance model was compared with two existing global population datasets, GPWv3 and LandScan 2005 and the Kenyan 1999 census data.

3.1 *Airborne remote-sensing data*

The 2004 digital camera data were acquired with 60% overlap and 30% sidelap between lines using a Nikon D1X colour digital camera equipped with a 14 mm lens producing a 78° opening angle. The camera is part of the EnsoMOSAIC system consisting of flight-planning software, navigation software, triggering unit, a global positioning system (GPS) and a power source (Holm *et al.* 1999). Data acquisition took place on 25 January between 8 and 9 am and on 27 January 2004 between 12 and 1 pm at altitudes between 2100 and 2700 m above the land surface, resulting in an approximate ground resolution between 27 and 40 cm for the study area. Brightness variations within the single frames of the 2004 data were removed by corrections for light-falloff effect and bi-directional effects using the methods developed by Pellikka (1998), after which the frames were mosaicked using EnsoMOSAIC (Holm *et al.* 1999). The resulting mosaics were orthorectified, projected to transverse mercator projection with a Clarke 1880 spheroid and Arc 1960 datum and resampled to 0.5 m ground resolution. The geometric accuracy was within 2 m as verified in the field using the GeoXT™ GPS (Trimble Navigation Ltd, Sunnyvale, CA, USA) with differential correction base (reference) data.

3.2 *Satellite remote-sensing data*

A SPOT 4 HRVIR 1 satellite image (15 October 2003, path & row 143-357, view angle 10.4°) with 20 m pixel size was used to derive surface-reflectance, image-texture and land-cover based predictors. The image was orthorectified utilizing a 20 m planimetric resolution digital elevation model (DEM), interpolated from 15.45 m interval contours captured from 1:50 000 scale topographic maps, then atmospherically corrected utilizing the historical empirical line method (HELM; Clark and Pellikka 2005). As noted by Moran *et al.* (2001), there is a near-linear relationship between at-satellite radiance and the surface-reflectance factor through the range 0–70% reflectance. Consequently, an accurate estimation of the HELM correction lines for each SPOT spectral band can be obtained using only two within-scene reflectance targets: firstly, a high-reflectance and spectrally invariant-in-time calibration site measured in the field and, secondly, an estimate of path radiance derived directly from the imagery

through identified within-scene dark objects. The HELM calibration target used was a vegetation-free roadside quarry, which was ~60 m wide and 200 m long, whilst the within-scene dark objects were areas of topographic shadow for the green and red bands and an area of standing water for the near-infrared (NIR). Surface-reflectance factor measurements were made during field work in the Taita Hills in 2005, using a FieldSpec® Handheld VNIR (325–1075 nm, 3.5 nm spectral resolution) spectroradiometer (ASD Inc., Boulder, CO, USA) calibrated to a Spectralon® reflectance reference panel before each measurement set. The spectrometer was handheld in a nadir view position at ~1.2 m height facing towards the sun, with a 25° bare-head optic giving a ground instantaneous field-of-view of 53 cm in diameter. In order to provide validation data for the HELM correction lines, measurements were also made of a sandy school playground, a compacted red-soil road area and an area of tarmac hard standing. Furthermore, a GPS was used to record the position of the centre of each of the targets so their location could be accurately determined within the SPOT image.

Finally, a topographic normalization was applied using band-specific '*c*' correction factors calculated for identified general vegetation classes (Teillet *et al.* 1982, Clark and Pellikka 2009). The satellite data acquisition date in October coincides with the end of the dry season. Before the short rains occurring in November to December, the croplands in the hills have just been planted (Jaetzold and Schmidt 1983) and are without vegetative cover and the biomass of the forests and woodlands is at its lowest. Consequently, differentiation between croplands and perennial vegetation is more easily accomplished than with rainy season imagery.

3.3 *Dwelling-unit data derived from airborne digital image mosaics*

The dependent variable dwelling units were interpreted by on-screen digitization from airborne digital image mosaics covering 30% of the Taita Hills using ArcGIS 9.3 software (Esri Inc., Redlands, CA, USA) (figure 1). To improve interpretation accuracy of dwelling units, ground-reference data were collected during January 2005 and 2006. A random non-stratified household survey ($n = 100$) carried out in October 2006 resulted in an average of 6 persons per dwelling unit, whereas a survey carried out by Soini (2005) resulted in an average of 6.2 people ($n = 45$). This number of inhabitants is somewhat higher than the average of 4.7 inhabitants for the Wundanyi division based on the population census of 1999. However, as the census presents fewer inhabitants per household for sub-locations in the lowlands, a reliable figure to represent the average number of dwellers per household for the whole of the Taita Hills was considered to be six people. This was the figure used with human population-abundance regression analysis.

3.4 *Satellite-image classification*

The SPOT image was classified into land-use and land-cover (LULC) types, according to nomenclature that was derived using the land-cover classification system (LCCS) protocol developed by the Food and Agriculture Organization (FAO) of the UN and the UN Environment Programme (Di Gregorio 2005). The LCCS software generates unique codes and Boolean formulas for each class, which allows other users to reconstruct the definitions used. The land-cover classes were defined based on the inspection of the airborne remote-sensing and SPOT data and fieldwork (Clark and Pellikka 2009).

A multi-scale segmentation/object relationship modelling (MSS/ORM) approach using Definiens eCognition software (Trimble Navigation Ltd, Sunnyvale, CA, USA) was applied for land-cover classification of the SPOT data (Burnett and Blaschke 2003, Baatz *et al.* 2004, Clark and Pellikka 2009). In a multi-scale segmentation, a so-called 'scale' parameter is used to determine the average size of the image segments at each level in the hierarchy. The first segmentation level is critical because the borders defined at this stage will be adhered to by any subsequent segmentations, either subdividing the image-object primitives or combining them into larger objects. In the small-scale cultivation areas in the hills, a very detailed initial segmentation with a scale parameter of two was used, whilst in the shrublands in the foothills and lowlands, an aggregation with a scale parameter of 4 was used. These segmentation levels were merged to the final mapping level 2. In the classification process itself, various segmented image-object spectral, contextual and hierarchical properties were used to determine the land-cover type of each image segment. The output map was subject to final visual inspection and manual editing of any noted errors. Ground-reference test data were collected during field work in January 2005 and 2006 using stratified random sampling and used in the classification accuracy assessment together with the airborne digital camera data acquired three months after the SPOT acquisition date. The overall accuracy of the final manually edited land-cover map was 89%, with a kappa index for agreement of 0.87. The class-specific producer's and user's accuracy assessment and image-segmentation methodology are discussed in more detailed in Clark and Pellikka (2009).

3.5 *Predictor variables*

We derived three sets of predictors from the THEMS database; a first set of remote-sensing-based variables: reflectance, texture, land cover; a second set of variables derived from geospatial datasets: elevation, aspect, slope, TWI, precipitation, irradiance and distance from roads or rivers; and a third set of combined variables. All predictors were organized in a geospatial database at 20 m grid resolution to match with the spatial resolution of the SPOT data. The *Zonal statistics* or *summarize* functions (ESRI 1991) in ArcGIS 9.3 was used to summarize mean, majority or percentage values within each of the 100 m analysis squares across the study area. The predictor variables are listed in table 1.

3.5.1 Predictors from remote-sensing data. The SPOT satellite data were used to derive the predictors from reflectance statistics, image texture and the classified land-cover map. First-order image statistics were the mean surface reflectance of SPOT bands 2 and 3 (red and NIR). We excluded band 1 (green) because of high Pearson's correlation (r) with the red band ($r > 0.95$). For second-order image-texture measurements, we used angular second moment, contrast, correlation, sum of squares variance, inverse difference moment and entropy, as they are the most relevant texture measures according to Baraldi and Parmiggiani (1995). Three different sizes of moving windows in Geographic Resources Analysis Support System (GRASS) 6.3 software (Open Source Geospatial Foundation, Vancouver, BC, Canada) were tested: 3×3, 7×7 and 15×15, with the result that 3×3 resulted in a 'salt and pepper' effect and the 15×15 window in a very strong smoothing effect. Consequently, we employed the 7×7 window size, as in Shaban and Dikshit (2001). The texture measures were calculated in four directions (0°, 45°, 90° and 135°) and averaged, as suggested by

Haralick *et al.* (1973). Second-order image-texture measures based on the GLCM were calculated for both the red and NIR band, but due to high correlation, only texture measures for the red band were used in the regression analysis. Furthermore, red-band second-order image-texture measures had strong correlation with each other, except for angular second moment and correlation ($r < 0.1$). Therefore, we accepted only angular second moment and correlation as second-order image-texture variables for regression analysis. To summarize image-texture measures in the 100 m analysis squares, we calculated the mean of pixel values from the texture images. Only four land-cover classes (croplands, thicket, woodland and plantation forest) were used from the LULC due to relatively low prevalence or high correlation with other predictors of the other land-cover classes. The percentage of spatial coverage for different land-cover classes in each 100 m analysis square was calculated using the *summarize* function in ArcGIS 9.3.

3.5.2 Predictors from geospatial datasets. A 20 m planimetric resolution raster DEM was interpolated from 15.45 m interval contour lines (captured from Survey of Kenya 1:50 000 scale topographic maps), utilizing the TOPOGRID function in ArcGIS, which is based upon the ANUDEM programme (Hutchinson 1989). The method applies a discretized thin plate spline technique, in which the roughness penalty has been modified to allow the fitted DEM to follow abrupt changes in relief, such as streams and ridges, which is useful in rugged terrain. The spot-height information on the scan maps was not used in the interpolation process, but rather these heights were used to assess an altimetric root mean square error (RMSE) for the DEM of ± 8 m, whilst a digitization accuracy of ± 1 mm derived a planimetric accuracy estimate of ± 50 m. The mean elevation, slope and aspect were calculated from the 20 m DEM and a commonly used indirect soil moisture measurement, TWI, was derived using a custom-made geoprocessing model in ArcGIS 9.3. The TWI (ω) was calculated using the following formula:

$$\omega = \ln (A_s / \tan \alpha) \tag{1}$$

where ln denotes the natural logarithm, A_s represents the upslope contributing area and α the slope angle (Beven and Kirkby 1979, Moore *et al.* 1991). Irradiance, solar radiation energy received on a given surface area in a given time (kW h m^{-2} month^{-1}), was calculated from the DEM using an ArcInfo AML macro (shortwavc.aml) (Esri Inc., Redlands, CA, USA) (Kumar *et al.* 1997, Zimmermann 2000). For the analysis, yearly mean irradiance values were scaled from 0 to 1. Long-term mean precipitation grids were interpolated from monthly available meteorological data and surrounding areas using ANUSPLIN software (Hutchinson 1995, Erdogan *et al*. 2011). Euclidean distance to main roads and main rivers, digitized from the Kenya 1:50 000 scale topographic map, were calculated using the *Euclidean distance* function in ArcGIS 9.3.

3.6 *Statistical analysis techniques*

We used GAMs (Hastie and Tibshirani 1990), which are a non-parametric extension of generalized linear models (GLMs) (McCullagh and Nelder 1989) to relate the dwelling units to remotely sensed and geospatial predictor datasets. GAMs were fitted using a logit link function and binomial error distribution for dwelling-unit distribution models and a Poisson error distribution via logarithmic link function for

dwelling-unit abundance models (Hastie and Tibshirani 1990). GAMs were run using smoothing splines with four degrees of freedom as default (Venables and Ripley 2002).

3.6.1 Model calibration. Models were calibrated using the GRASP 3.3 package (Lehmann *et al.* 2002) for S-PLUS 6.2 (TIBCO Software Inc., Somerville, MA, USA). The full dataset for model building ($n = 10\,488$) was randomly divided into a calibration dataset including 70% ($n = 7342$) of the samples and into an evaluation dataset ($n = 3146$) using random selection in SPSS 15.0 for Windows (IBM Corporation, Armonk, NY, USA) following the split-sample approach (Guisan and Zimmermann 2000).

For three different datasets (full, calibration and evaluation data), we fitted three types of regression models: two partial models and one full model, which enabled us to study the individual model performance. The first partial model used only the remote-sensing-based predictors, the second only the geospatial predictors and the full model all the predictors. In each modelling case, we started from a complete model with all variables included. Stepwise regression procedures were used in model selection based on Akaike's information criterion (AIC) (Akaike 1974).

3.6.2 Model evaluation. All models were evaluated as follows:

1. By using the percentage of explained deviance as an indicator for model explanatory power (D^2). It was obtained by dividing the difference between null and residual deviance by the null deviance. We adjusted the D^2 value following Weisberg (1980) and Guisan and Zimmerman (2000) as:

$$\text{adj.}D^2 = 1 - [(n-1)/(n-p)][1 - D^2] \tag{2}$$

   The value of the adjusted D^2 (adj.D^2) increases with an increasing number of observations (n) or a decreasing number of parameters (p) in the model. This approach corrects the D^2 (deviance explained) for the number of fitted regression parameters and the number of observations, thus considering the degrees of freedom.
2. By using the area under the curve (AUC) from the receiver operating characteristic plot to indicate the model predictive power (Fielding and Bell 1997). A rough guide for classifying the accuracy of the AUC is the traditional academic point system (Swets 1988): 0.90–1.00 = excellent; 0.80–0.90 = good; 0.70–0.80 = fair; 0.60–0.70 = poor; 0.50–0.60 = fail.
3. By using Cohen's kappa statistic (Cohen 1960). The kappa value was calculated using optimal thresholds determined with the *PresenceAbsence* R-package (http://cran.r-project.org). kappa scores were calculated for 100 threshold values (in 0.01 increments), and the one that provided the highest kappa became the accepted threshold (Segurado and Araújo 2004). According to Landis and Koch (1977), models can be classified based on the kappa statistics into: poor, $\kappa < 0.00$; slight, $\kappa = 0.00$–0.20; fair, $\kappa = 0.21$–0.40; moderate, $\kappa = 0.41$–0.60; substantial, $\kappa = 0.61$–0.80 and almost perfect, $\kappa = 0.81$–1.00.
4. By calculating the contribution for each predictor in the combined distribution and abundance models using a calibration dataset. Model contribution gives an indication of the contribution of the variable within the selected model.

It corresponds to the possible range of variation on the scale of the linear predictor (Lehmann *et al*. 2002).

5. Predictor maps from selected models were used to predict probability of dwelling occurrence on a 100 m grid resolution for the whole study area. Prediction maps were built in ArcView 3.2 (Esri Inc., Redlands, CA, USA). Models were first exported from S-PLUS as lookup tables and processed in ArcView 3.2 by a ready-made Avenue script, which is a part of the GRASP package. This script reclassifies the predictors maps corresponding to those selected in the model (Lehmann *et al.* 2002).

3.7 *Comparison of model output with global and Kenyan population data*

LandScan 2005 and GPWv3 global population grids covering the study area were converted from grid format to vector format as points using ArcGIS 9.3. Each point then represented the total population of the grid cell. As a result, we were able to compare our predictive population models with LandScan 2005 and GPWv3 datasets by selecting points that fell inside the Taita Hills study area. We also made upscaling operations in ArcGIS 9.3 by summing the LandScan 1 km population data to match with the GPWv3 5 km grid and then doing the same with the 100 m population-abundance modelling data. This upscaling method proved to have a weakness, since some 5 km grids fell partly outside the study area (figure 2). In the end, only 14 grids were used for comparing the population. Correlation analysis was made between these three different datasets.

In order to compare predicted population models with Kenyan census data for 1999, we summed 100 m population-abundance modelling points that fell inside a specific sub-location polygon. The first dataset included the 32 sub-locations located totally above the 1100 m elevation zone and the second one included sub-locations extending also to the lowlands, totalling to 50 sub-locations (figure 2). Correlation between Kenyan census data for 1999 and different population models was calculated for both datasets.

4. Results

4.1 *Explorative data analysis*

Before the model calibration process, correlations between predictor variables were investigated in order to avoid problems caused by multi-collinearity in regression analysis (Farrar and Glauber 1967). Despite high correlation ($r = -0.88$) between slope and irradiance, we kept both variables for modelling because of our interest in dependency of dwelling-unit locations on them. The correlation coefficient for all the other variables was lower. As an example, relatively high correlation ($r = 0.39$) occurred between reflectance values in the NIR band and elevation, expressing that there is more strongly reflecting green vegetation in higher areas, evidently as a result of the high rainfall and cooler temperatures (table 2).

4.2 *Explanatory and predictive power of the models*

The distribution models explained 19–31% of variation in the dwelling-unit occurrence data (adj.D^2). When comparing the explanatory power for calibration and

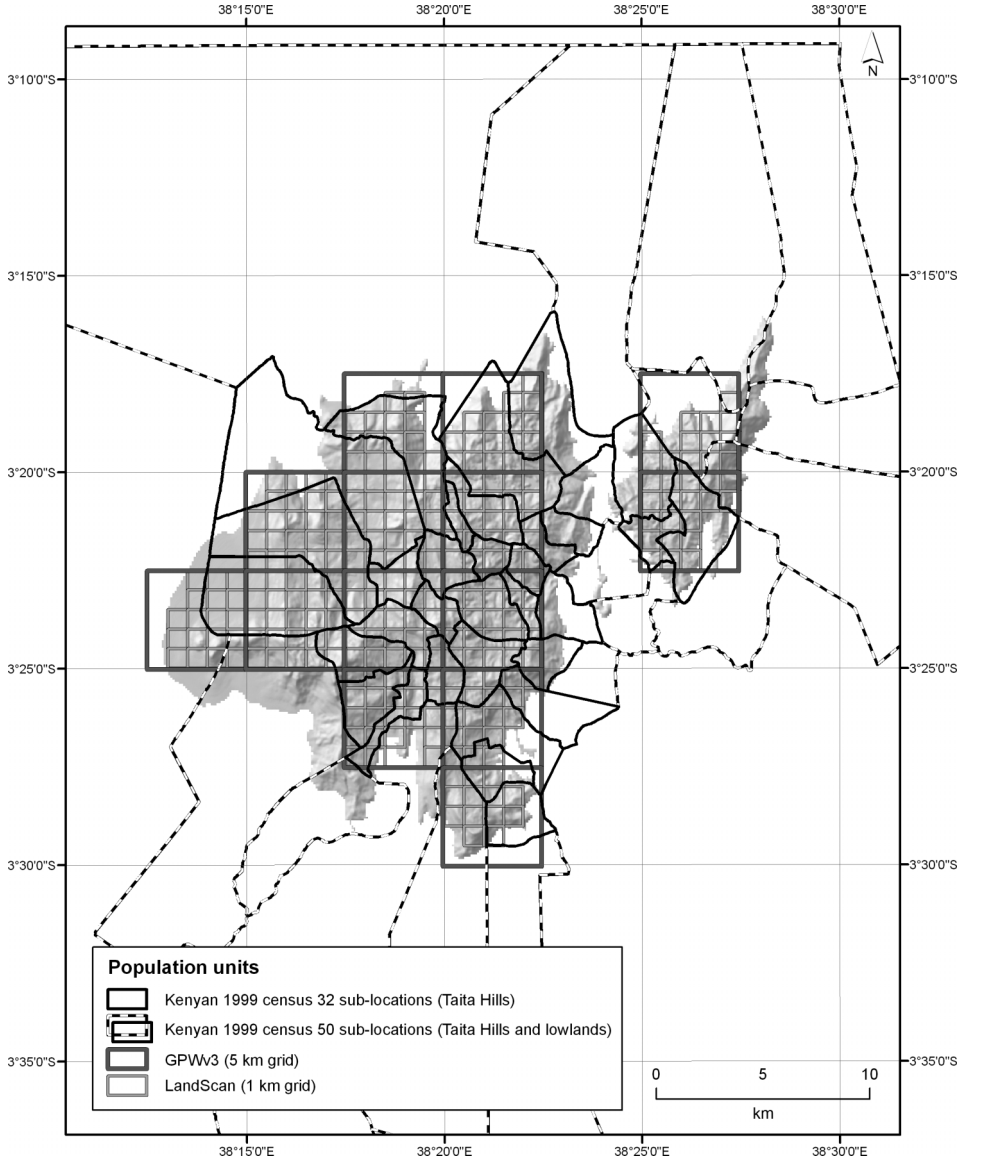

Figure 2. Study design for comparing different population datasets: LandScan 2005 (1 km); GPWv3 (5 km); Kenyan census data for 1999 sub-locations in the Taita Hills ($n = 32$) and sub-locations also extending into the lowlands ($n = 50$). GAM model grids (100 m) not shown.

evaluation data, the adj.D^2 was only slightly lower for evaluation data, confirming that the models were stable. Predictors from remote-sensing data explained 25% of the dwelling-unit distribution, whilst predictors derived from other geospatial data explained 20%. Combined models with all predictors had the best model performance, explaining 31% of the variation in dwelling-unit data distribution (table 3).

Table 2. Correlation matrix of the environmental predictors used in the statistical analyses. Correlation coefficients and statistical significances were derived from the bivariate correlation procedure.

Variable	Elevation	Aspect	Slope	TWI	Irradiance	Precipitation	Distriver	Distroad	red band	NIR band	Asm2	Corr2	Croplands	Thicket	Woodland	Plantation forest
Elevation		**	**	**	**	**	**	**	**	**	n.s.	**	**	**	**	**
Aspect	−0.14		n.s.	**	**	**	*	**	**	**	**	**	**	**	**	*
Slope	0.17	0		**	**	**	**	**	**	**	**	**	**	**	**	*
TWI	−0.1	0.04	−0.09		n.s.	**	**	**	**	**	**	**	n.s.	n.s.	**	**
Irradiance	−0.09	−0.02	−0.88	0.01		**	**	**	**	**	**	**	**	**	**	**
Precipitation	0.61	−0.2	0.25	−0.44	−0.18		**	**	**	**	**	**	**	**	**	**
Distriver	0.04	0.01	0.26	−0.62	−0.15	0.49		**	**	**	**	**	**	**	**	**
Distroad	−0.23	0.12	0.21	0.08	−0.19	−0.14	0.12		**	**	**	**	**	**	n.s.	**
red band	−0.22	0.06	−0.35	0.14	0.3	−0.47	−0.18	−0.03		**	**	**	**	**	**	**
NIR band	0.39	−0.14	0.04	0.02	−0.09	0.4	−0.12	−0.28	−0.28		**	*	**	**	**	**
Asm2	0	0.06	0.14	−0.03	−0.13	0.16	0.13	0.27	−0.4	−0.05		n.s.	**	**	**	**
Corr2	0.09	−0.04	0.1	−0.04	−0.08	0.03	0.03	−0.06	−0.01	0	−0.4		**	**	**	**
Croplands	0.22	−0.21	−0.22	0.01	0.2	0.13	−0.13	−0.38	0.4	0.36	−0.38	0.02		**	**	**
Thicket	−0.27	0.07	0.2	0	−0.12	−0.22	0.11	0.31	−0.05	−0.58	0.11	0.03	−0.46		**	**
Woodland	0.09	−0.05	0.2	0.02	−0.21	0.25	0.02	0.01	−0.52	0.51	0.06	0.07	−0.17	−0.25		**
Plantation forest	0.26	−0.01	0.01	−0.2	0.01	0.33	0.1	−0.14	−0.44	0.06	0.11	0.1	−0.17	−0.2	−0.04	

Note: Pearson's correlation (** = $p < 0.01$, * = $p < 0.05$, n.s. = non-significant).

Table 3. The explanatory power (deviance explained) of model fit. Adjusted D^2 values are listed for the models containing both predictor sets (combined), the geospatial predictors only (geospatial models), and the remote-sensing-based (RS) predictors, respectively.

	Geospatial models adj.(D^2)	RS models adj.(D^2)	Combined models adj.(D^2)
Distribution models			
All data ($n = 10\,488$)	0.20	0.24	0.30
Calibration data ($n = 7342$)	0.20	0.25	0.31
Evaluation data ($n = 3146$)	0.19	0.24	0.29
Abundance models			
All data ($n = 10\,488$)	0.30	0.38	0.46
Calibration data ($n = 7342$)	0.32	0.39	0.47
Evaluation data ($n = 3146$)	0.28	0.36	0.44

Fourteen variables with smoothing splines and four degrees of freedom were included in the final GAM model for population distribution. The regression formula has the form:

Dwelling-unit presence $\sim s$(Elevation, 4) + s(Aspect, 4) + s(Slope, 4) + s(TWI, 4) + s(Precipitation, 4) + s(Distriver, 4) + s(Distroad, 4) + s(red band, 4) + s(NIR band, 4) + s(Asm^2, 4) + s(Corr2,4) + s(Croplands, 4) + s(Thicket, 4) + s(Woodland, 4) where s = spline smoother and 4 is the number of degrees of freedom for the spline smoother.

All distribution models were capable of discriminating between presence and absence dwelling units, with AUC values ranging from 0.80 for models based on geospatial data to 0.86 for combined models. According to the classification by Swets (1988), discrimination is low if AUC < 0.7, fair if $0.7 <$ AUC < 0.8, good if $0.8 <$ AUC < 0.9 and excellent if AUC > 0.9. Consequently, all the models had good discrimination capacity (table 4).

Abundance models explained 28–47% of the variation in dwelling-unit abundance data (adj.D^2). Models based on the geospatial data explained c. 28% to 30%, models based on the remote-sensing data explained 36–39% and models combining both sources explained 44–47% of the variation in dwelling-unit abundance (table 3). Fifteen variables with smoothing splines and four degrees of freedom were included in the final GAM model for population abundance. The regression formula has the form:

Population abundance $\sim s$(Elevation, 4) + s(Aspect, 4) + s(Slope, 4) + s(TWI, 4) + s(Precipitation, 4) + s(Distriver, 4) + s(Distroad, 4) + s(red band, 4) + s(NIR band, 4) + s(Asm2, 4) + s(Corr2, 4) + s(Croplands, 4) + s(Thicket, 4) + s(Woodland, 4) + s(Plantation forest, 4) where s = spline smoother and 4 is the number of degrees of freedom for the spline smoother.

Table 4. Model performance predictive power (area under the receiver operator curve, AUC).

Distribution models	Geospatial models AUC	RS models AUC	Combined models AUC
All data ($n = 10\,488$)	0.80	0.83	0.86
Calibration data ($n = 7342$)	0.80	0.83	0.86
Evaluation data ($n = 3146$)	0.80	0.82	0.85

Table 5. Model accuracy assessed using Cohen's kappa value (Cohen 1960).

Distribution models	Geospatial models (kappa)	RS models (kappa)	Combined models (kappa)
All data ($n = 10\,488$)	0.37	0.42	0.47
Calibration data ($n = 7342$)	0.37	0.42	0.47
Evaluation data ($n = 3146$)	0.36	0.41	0.46

Kappa statistics for distribution models were calculated using optimized thresholds in the *PresenceAbsence* R-package. The kappa value was the best for combined models, good for models based on remote-sensing data, but less satisfactory for models based on geospatial data (table 5).

4.3 *Model contributions of predictors*

Table 6 shows the contribution of each predictor within the selected population distribution and abundance models. Angular second moment image-texture measurements for SPOT red band (Asm2) had the highest contributions in both of the models. Other variables contributed less, but precipitation, mean elevation and reflectance in red and NIR bands played an important role in the distribution models. For abundance models, remote-sensing-based predictors also contributed the main part. Precipitation, distance to roads (Distroad) and elevation were the major geospatial contributors for abundance models.

4.4 *Partial response curves*

One of the key parts of the interpretation of GAM models is the description of the predictors' partial response curves represented in figure 3. The model predicts that

Table 6. Contribution in percentage of each predictor within the selected models (model contribution in GRASP).

	Contributions (%)	
Predictor	Distribution models	Abundance models
Elevation	3.73	2.19
Aspect	0.37	0.16
Slope	1.77	1.79
TWI	1.59	1.09
Irradiance	Not in the model	Not in the model
Precipitation	3.92	2.72
Distriver	0.96	0.72
Distroad	2.47	2.26
red band	3.57	2.49
NIR band	3.7	2.86
Asm2	14.9	13.36
Corr2	2.53	1.13
Crops	1.07	0.76
Thicket	0.64	0.67
Woodland	0.79	1.07
Plantation forest	Not in the model	0.61

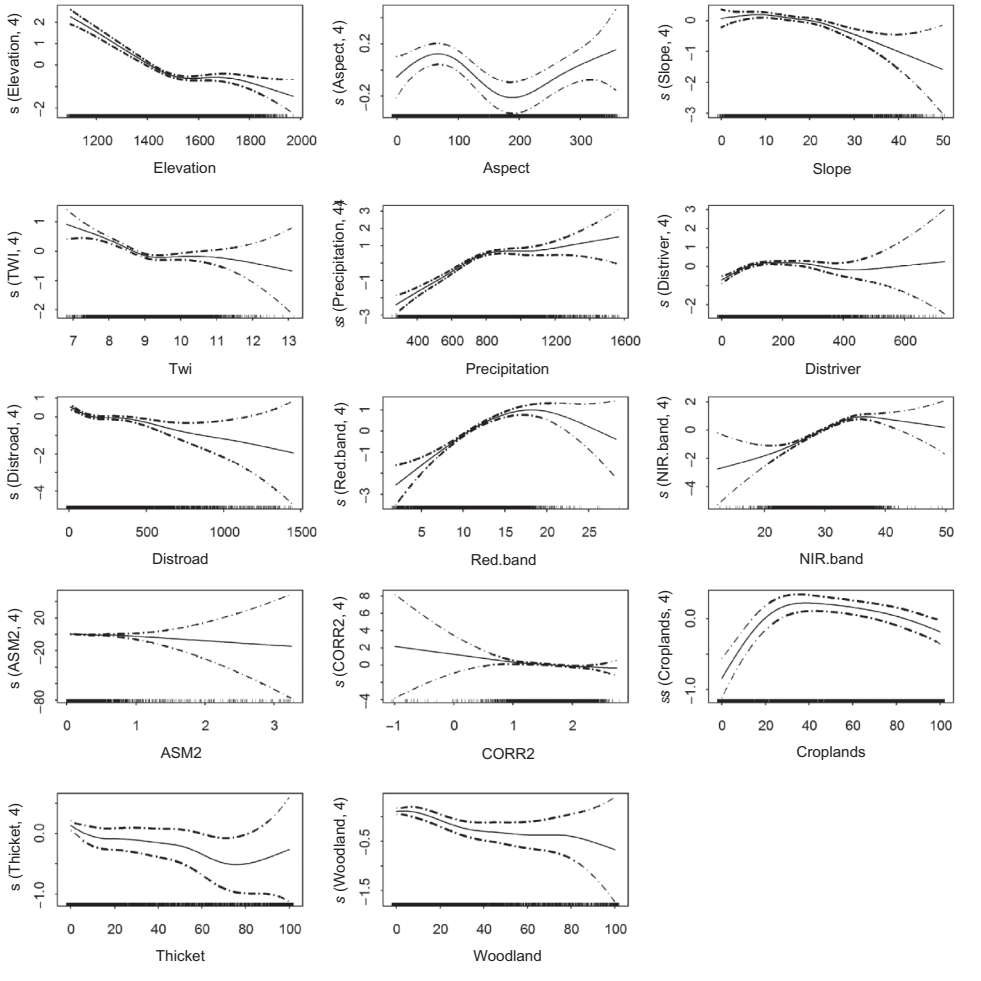

Figure 3. Response curves from the final population distribution GAM model based on Akaike's information criterion (AIC) selection method for dwelling units. Results are expressed in the scale of the additive predictor before transformation into the prediction scale by the inverse link function. Dashed lines represent upper and lower pointwise twice-standard-error curves. Tickmarks show the location of observations along the variable range. For definitions of the predictor variables units, see table 1.

more people tend to live at lower elevations between 1100 and 1300 m a.s.l. and on east and southeast facing slopes. This may be explained by rain-shadow effects related to the aspect causing the western slopes to be less favourable for agriculture. The response curve for slope shows that dwelling units are present on slopes less than 35°, whilst the maximum response was obtained at slope angles of 10°. Actually, almost all the land area above 1100 m a.s.l. in the Taita Hills is inclined. Fewer people are expected to live in very moist areas (TWI in figure 3), and there is a clear pattern that there are more houses in areas with higher precipitation. People seem not to live very close to rivers, since the maximum value being reached was at a distance of 200 m from rivers. The reason might be that in rough mountainous areas, the rivers are at the bottom of steep valleys, whilst the small brooks and springs attracting people are not

included in the river data derived from the topographic map. As expected, people tend to live close to roads. There is a curvilinear association between presence of dwelling units and red reflectance, with dwelling-unit occurrence peaking at a reflectance value of 17%, which may be explained by the fact that there is strongly reflecting barren land associated with built-up areas. The NIR band peaked at the reflectance value 35%, presumably caused by strongly reflecting orchard trees and croplands adjacent to dwelling units. Image-texture measurement angular second moment for the red band (Asm2) had more dwellings at low values, and the same type of almost linear decline can be seen for image-texture correlation measurement (Corr2). The response curve for the croplands had a positive association peaking at c. 40% and declining from there on, which is as expected since farming in the area is typically practised by small households. Dwellings were present at low values for thicket and woodland because agriculture is not favourable in those land-cover types. Irradiance was dropped out from the final distribution and abundance models because of the high correlation with slope variable, and plantation forest class was also dropped out from the distribution model due to the relatively low prevalence of the class in the dataset.

4.5 *Predicted human population distribution and abundance maps*

The GAMs were extrapolated for the whole study area, and dwelling-unit probability maps were produced using GIS techniques as shown in figure 4, in which it is seen that the model is capable of discriminating between inhabited and uninhabited areas. In figure 4, there are no dwelling units in Ngangao forest or in the lake and swamp in the northeastern part of the image, for example. The model is also capable of predicting human population concentrations in and around the villages and distinguishes the absence of dwellings on cultivated fields by giving a low dwelling-unit probability. A probability map of human population distribution extrapolated for the whole study area (Taita Hills above 1100 m) can be seen in figure 5.

4.6 *Model comparison with two existing population databases and Kenyan census data*

We compared our population-abundance models with two existing global scale population products, GPWv3 and LandScan 2005. There was a statistically significant correlation between our combined and remote-sensing-based models and the GPWv3 product ($r > 0.8$), but the correlation was non-significant with the geospatial model ($r = 0.19$). For LandScan 2005, the correlations were lower (table 7).

The correlation between Kenyan census data for 1999 and predicted population-abundance models are high for remote-sensing data ($r = 0.71$) and combined models ($r = 0.51$) when solely sub-locations over 1100 m a.s.l. ($n = 32$) were used. For geospatial models, the correlation was non-significant. There was low correlation ($r = 0.34$) between remotely sensed population-abundance models and Kenyan census data for 1999 for the sub-locations also extending into the lowlands ($n = 50$), and no correlation for combined and geospatial models (table 8).

5. Discussion

To give an answer to two simple questions, 'how many are we' and 'where do we live' is not an easy task, since estimation of human population distribution and abundance

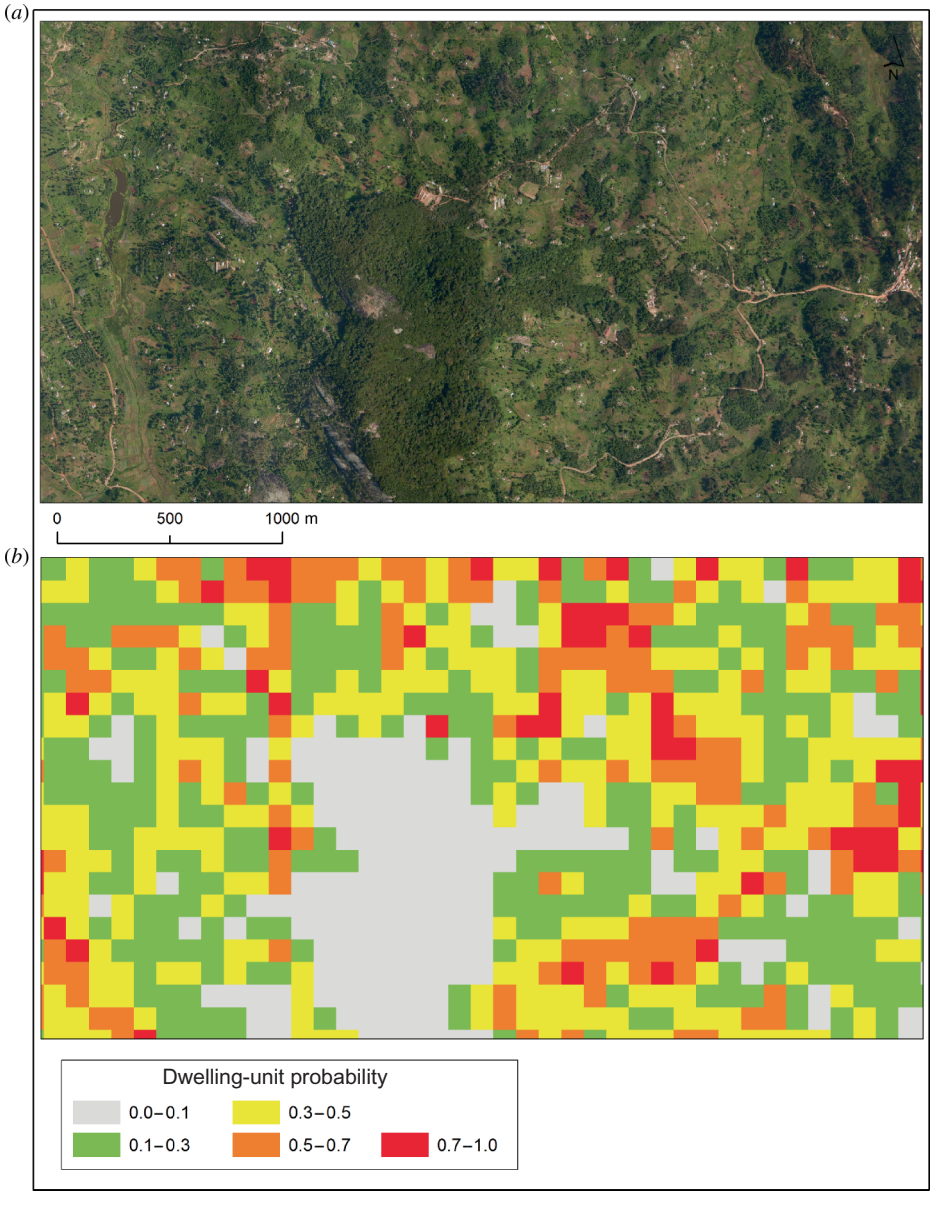

Figure 4. (*a*) Typical rural area in the Taita Hills around Ngangao indigenous forest in the digital camera mosaic acquired on 25 January 2004 and (*b*) the prediction map of inhabited areas around the same area (100 m grid). The location of Ngangao forest can be seen in figure 1.

is very challenging. In Europe, for instance, population census counts are carefully collected and stored in digital format, but in Kenya, for example, population data are gathered by counting people by a traditional census, which is a complex nationwide error-prone operation. The census is carried out at a household level, and the data are then aggregated to administrative units, from sub-location to district levels. Population

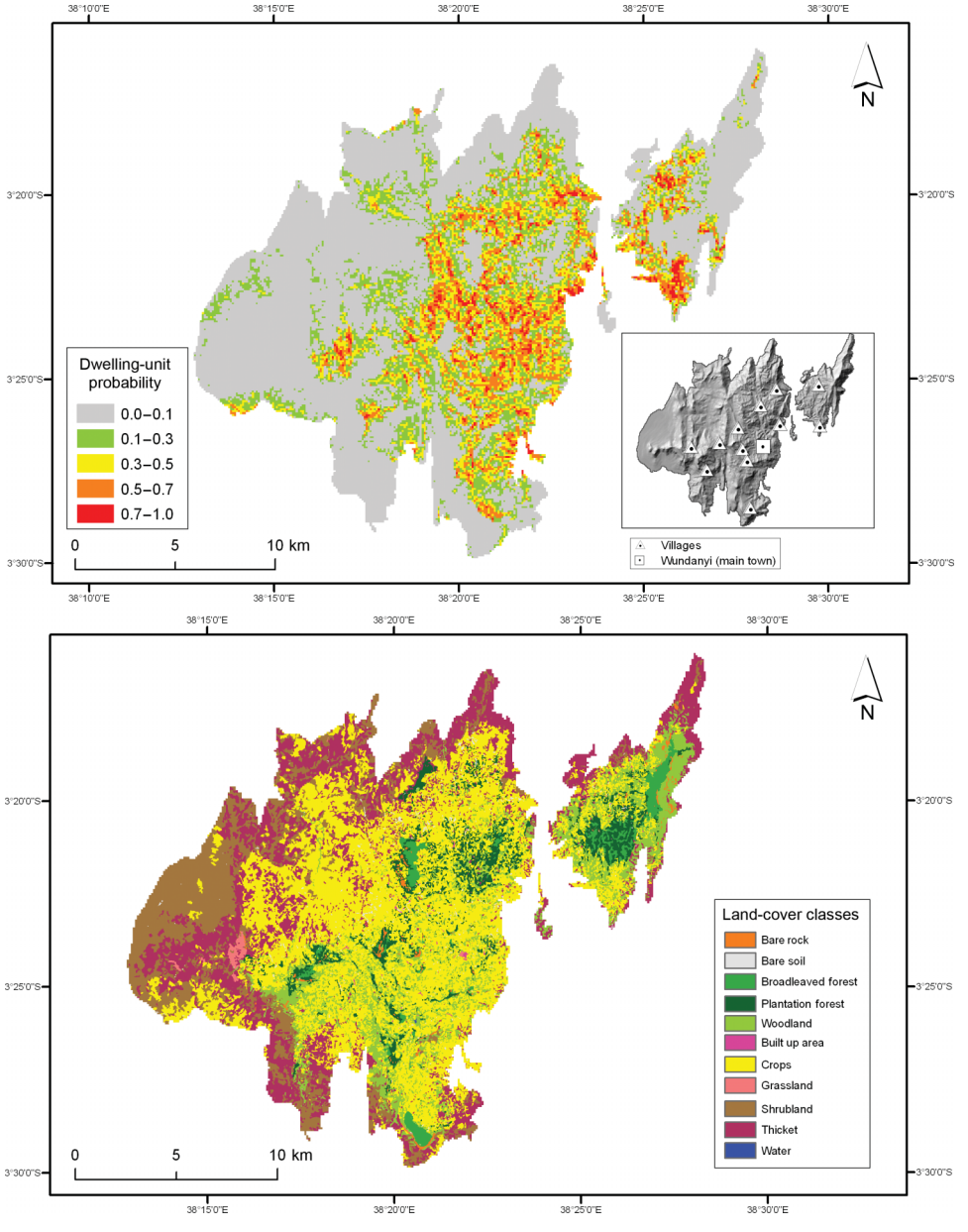

Figure 5. A map of land cover and the predicted inhabited areas in the Taita Hills. The probabilities for dwelling-unit occurrence were calculated for the whole study area with a calibrated GAM at 100 m resolution using combined predictors.

data in Kenya are determined by these administrative areas, even though often standardized units, such as regular analysis grids, are preferred in scientific applications. We used GRASP regression modelling based on airborne imagery and geospatial data for predicting human population distribution and abundance in the Taita Hills, Kenya, to obtain population estimates in regular analysis grids.

Table 7. Correlation between different population data sets upscaled to 5 km grid ($n = 14$). Correlation coefficients and statistical significances were derived from the bivariate correlation procedure.

Population model	GPW	LandScan	Combined (mean)	RS (mean)	Geospatial (mean)
GPWv3		n.s.	**	**	n.s.
LandScan 2005	0.47		*	**	n.s.
Combined (mean)	0.82	0.67		**	n.s.
RS mean	0.83	0.70	0.91		n.s.
Geospatial (mean)	0.19	0.20	0.53	0.19	

Note: Pearson's correlation (** = $p < 0.01$, * = $p < 0.05$, n.s. = non-significant).

Table 8. Correlation between population models and Kenya population census 1999 for sub-locations solely in the Taita Hills ($n = 32$) and for sub-locations also extending to lowlands ($n = 50$). Predicted population was summed to a specific sub-location. Correlation coefficients and statistical significances were derived from the bivariate correlation procedure.

	Taita Hills data ($n = 32$)			Taita Hills and lowland data ($n = 50$)		
	Combined	RS	Geospatial	Combined	RS	Geospatial
Kenya census 1999	0.54**	0.71**	0.11 n.s.	0.19 n.s.	0.34*	−0.07 n.s.

Note: Pearson's correlation (** = $p < 0.01$, * = $p < 0.05$, n.s. = non-significant).

5.1 *Model performance*

The results of the GAM models presented here suggest that the distribution and abundance of population can be explained using remote-sensing and geospatial data. The explanatory power of the models was moderate, and the predictive power was good. Predictors combined from remote-sensing and geospatial data had the best model performance, indicating that both variable groups are needed in modelling, which is in line with other studies. Lo (1995), for example, integrated digital imagery and geospatial data for population density regression analysis for the Hong Kong metropolitan area. Studies in species population distribution modelling, including remotely sensed variables and topographic and climatic variables (Thuiller *et al.* 2004, Zimmermann *et al.* 2007), confirmed that remotely sensed variables improved the fit of the models. Thuiller *et al.* (2004) found that the inclusion of land cover significantly improved the explanatory power of bioclimatic species models, whereas Zimmermann *et al.* (2007) analysed the partial contributions of remotely sensed and topographic–climatic predictor sets and concluded that the model fit was highest when using combined predictors. Buermann *et al.* (2008) found that distribution models that included topographic, climatic and remote-sensing-derived vegetation variables were more accurate than models including only topographic and climatic variables.

Remotely sensed variables alone have also proven to be good predictors in species population distribution studies. Zimmermann *et al.* (2007) found that models derived only with remotely sensed predictors had only slightly less satisfactory model performance than combined models. This was in accordance with our findings where the model explanatory and prediction performance was lower for only remotely sensed

variable groups than for combined models, but still the model fit was reasonably good. We used first- and second-order texture measurements, and in the final population distribution and abundance models, the angular second moment was the best predictor. St-Louis *et al.* (2006) also used image-texture measurements for species-distribution models with high model performance. Li and Weng (2005) used Landsat ETM+ imagery to estimate urban population density, and they found that integration of texture, temperature and spectral predictors substantially improved the accuracy of population estimation.

The normalized difference vegetation index (NDVI) is commonly used for assessing landscape characteristics in species modelling, even though it has significant correlation with the red and NIR bands, as it is calculated from these reflectances. We investigated NDVI values for human population models, but eliminated it due to the high correlation with the red and NIR bands. A relatively low performance of land-cover variables made us wonder if it was necessary to include land-cover classification at all, given the production effort needed to generate such data, when the first-order image statistics and second-order texture measurements from the original SPOT imagery had higher contributions to the population models. Guisan and Zimmermann (2000) and Bellis *et al.* (2008), for example, stated that land-cover maps derived from remote sensing are often not detailed enough to improve predictions of species distribution models.

Variables derived from geospatial data had by far the lowest model performance in our population models, and therefore we suggest that they should not be used alone as predictors for dwelling-unit distribution and abundance modelling. However, availability of a more precise road network and hydrographic network might have improved the model performance. For example, in the rural areas of the Taita Hills, the majority of houses are accessed only by footpaths, which are not included in the road network (which we digitized from 1:50 000 scale maps). Therefore, the distance to roads resulted only in an indicative result. Other predictors from geospatial data, for example, soil type, could also be tested for model improvement. Precipitation was the main determinant factor derived from geospatial data for human population distribution at a local scale. Due to orographic rainfall patterns, rainfall is more abundant in higher elevations, especially on the south and southeastern slopes. Another reason why elevation correlated with abundance of dwelling units is lower temperatures at higher elevations causing decreased evapotranspiration for plants and a more tolerable climate for people. However, temperature was left out from the modelling due to an almost linear correlation with elevation. In our final models, distance-based variables, especially the distance to roads variable, had a surprisingly low role, as discussed earlier.

When calculating the GRASP model contribution (table 6), the angular second moment image-texture measurement had the highest contribution. Second-order image-texture factors have been shown to be important factors in urban population-density analysis (Shaban and Dikshit 2001, Li and Weng 2005), and species-occurrence analysis (St-Louis *et al.* 2006). These results are in accordance with our study, where both texture measurements, angular second moment and correlation had an important contribution in dwelling-unit modelling performance. First-order image statistics values, that is, red and NIR reflectance from the SPOT data, were also important predictors.

There was statistically significant correlation between our combined model and the model derived from remote-sensing data and the GPWv3 product, but the correlation

was non-significant, with the model derived from geospatial data. For LandScan 2005, the correlations were lower, which indicates that GPWv3 should be used instead of the LandScan product if population is estimated for rural mountainous areas in Africa. However, the coarse resolution of 5 km of GPWv3 at the equator makes it a bit less usable than the LandScan product, which has a resolution of 1 km. It can be concluded that GPWv3 is too coarse for local-scale population studies, and the accuracy of LandScan 2005 is questionable. Having stated this, it needs to be borne in mind that population models are place and time dependent, and our population-abundance model for the Taita Hills relates solely to dwelling-unit count and is based on the assumption that six persons live in each dwelling unit. In reality, the exact numbers might vary though.

Local-scale predictive human population models proved to have high correlation with the Kenyan census data for 1999 in the Taita Hills, but low or no correlation at all when the sub-locations that extended into the lowlands were analysed. This finding confirms that the present sub-location units are inadequate to model the population distribution in mountainous rural areas in which the population density is highly variable within the sub-location. The difference between our population prediction in the area over 1100 m a.s.l. and the census could also be partially explained by the population growth and the time gap between 1999 and 2005 and 2006, when the field survey was carried out.

In this study, we omitted socio-economic predictors, except distances calculated to the main roads, even though socio-economic factors, such as kinship, cost of land, land tenure and migration patterns, are important determinants of human settlement patterns. The reason these factors were omitted is that they are often very hard or even impossible to collect and use in a proper manner in predictive human population grid-based geospatial analyses, especially in studies conducted in data impoverished developing countries. This study deliberately aimed at testing the suitability of using solely remote-sensing and geospatial predictors to estimate human population distribution and abundance at a local scale. Based on the results of this study, we believe that the omission of socio-economic factors may affect modelling performance. This factor probably manifested itself in our models as the fairly low predictive power in explaining 19–31% of variation in the dwelling-unit occurrence data and 28–47% of the variation in dwelling-unit abundance data. Therefore, we recommend the inclusion of socio-economic factors in human prediction models if such data have been collected for the study area in question and are available in a suitable way for predictive modelling. However, we believe that the results of this study in estimating population distribution and abundance using solely remote-sensing and geospatial predictors are encouraging. Therefore, we suggest that these models could be used if up-to-date census data is not available or if the resolution of existing grid-based population models is too coarse for the study purpose. Predictive modelling techniques can be considered as a noteworthy alternative for human population distribution and abundance analysis, especially in areas of rapid population growth and land-cover change, such as Africa.

6. Conclusions

Accurate data of human population size and distribution are not available for many parts of the world or are of poor quality. In a local-scale population study, the global population products GPWv3 and LandScan 2005 proved to be ill-fitting to estimate the Taita Hills population at a fine scale, and population census data, based on

sub-locations as geographical and statistical units, proved to be cumbersome. The predictive models using predictors from remote-sensing and geospatial data created here were found to be more accurate than global datasets and correlated well with the census data too. However, it must be kept in mind that the modelling can often be applied only with sufficient datasets within a limited geographical extent.

Consideration should be given to the possibilities of local specialists to replicate the methodology in the creation of population and other kinds of geospatial models for Africa. Models should be straightforward and the GIS and statistical software used should be free, low-cost or open-source software. In this study, we used GRASP with commercial software S-PLUS 6.2 (Insightful Corp.), but there is also a GRASP package for the freeware R-program (R Development Core Team 2008). Similarly, low-cost or free GIS software (e.g. GRASS) could be used for the creation of geospatial datasets. Finally, based on our experience, we highly recommend that the GRASP method should be used in manifold predictive modelling for the sustainable management of the vulnerable environment.

Acknowledgements

The research was financially supported by the Academy of Finland funded TAITATOO project (http://www.helsinki.fi/science/taita). Thanks to Hanna Piepponen, who assisted in interpreting and digitizing the dwelling-unit points and to Milla Lötjönen and Antero Keskinen, who created the airborne digital image mosaics. Two anonymous reviewers greatly improved an earlier version of the manuscript.

References

Akaike, H., 1974, A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **19**, pp. 716–722.

Baatz, M., Benz, U., Dehghani, S., Heynen, M., Höltje, A., Hofmann, P., Lingenfelder, I., Mimler, M., Sohlbach, M., Weber, M. and Willhauck, G., 2004, *eCognition Professional: User Guide 4* (Munich, Germany: Definiens Imaging).

Balk, D. and Yetman, G., 2004, *The Global Distribution of Population: Evaluating the Gains in Resolution Refinement, Documentation for GPWv3* (Palisades, NY: CIESIN, Columbia University). Available online at: http://beta.sedac.ciesin.columbia.edu/gpw/documentation.jsp (accessed 19 November 2009).

Baraldi, A. and Parmiggiani, F., 1995, An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters. *IEEE Transactions on Geoscience and Remote Sensing*, **33**, pp. 293–304.

Bellis, L.M., Pidgeon, A.M., Radeloff, C.V., St-Louis, V., Navarro, J.L. and Martella, M.B., 2008, Modeling habitat suitability for greater rheas based on satellite image texture. *Ecological Applications*, **18**, pp. 1956–1966.

Beven, K.J. and Kirkby M.J., 1979, A physically based, variable contributing area model of basin hydrology. *Hydrological Science Bulletin*, **24**, pp. 43–69.

Buermann, W., Saatchi, S., Smith, T., Zutta, B., Chaves, J., Mila, B. and Graham, C., 2008, Predicting species distributions across the Amazonian and Andean regions using remote sensing data. *Journal of Biogeography*, **35**, pp. 1160–1159.

Burnett, C. and Blaschke, T., 2003, A multi-scale segmentation/object relationship modelling methodology for landscape analysis. *Ecological Modelling*, **168**, pp. 233–249.

Center For International Earth Science Information Network (CIESIN), Columbia University, and Centro Internacional De Agricultura Tropical (CIAT), 2005, *Gridded Population of the World Version 3 (GPWv3): Population Grids* (Palisades, NY:

Socioeconomic Data and Applications Center (SEDAC), Columbia University). Available online at: http://sedac.ciesin.columbia.edu/gpw (accessed 19 November 2009).

Clark, B.J.F. and Pellikka, P.K.E., 2005, The development of a land use change detection methodology for mapping the Taita Hills, South-East Kenya. In *Proceedings of the 31st International Symposium of Remote Sensing of the Environment*, 20–24 June 2005, St Petersburg, Russia. CD-ROM.

Clark, B.J.F. and Pellikka, P.K.E., 2009, Landscape analysis using multiscale segmentation and object orientated classification. In *Recent Advances in Remote Sensing and Geoinformation Processing for Land Degradation Assessment*, A. Röder and J. Hill (Eds), pp. 323–342 (London: Taylor & Francis).

Cohen, J., 1960, A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, pp. 37–46.

Di Gregorio, A., 2005, *Land Cover Classification System (LCCS)*, *Version 2: Classification Concepts and User Manual*, Food and Agriculture Organization (FAO) Environmental and Natural Resources Series 8, pp. 208 (Rome, Italy: FAO).

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C. and Worley, B.A., 2000, LandScan: a global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*, **66**, pp. 849–857.

Environment Systems Research Institute (ESRI), 1991, *ARC/INFO User's guide. Cell-based modelling with GRID. Analysis*, *Display and Management* (Redlands, CA: ESRI).

Erdogan, E.H., Pellikka, P. and Clark, B., 2011, Modelling the impact of land-cover change on potential soil loss in the Taita Hills, Kenya, between 1987 and 2003 using remote-sensing and geospatial data. *International Journal of Remote Sensing*, to be published, doi: 10.1080/01431161.2010.499379.

Farrar, D.E. and Glauber, R.R., 1967, Multicollinearity in regression analysis: the problem revisited. *Review of Economics and Statistics*, **49**, pp. 92–107.

Fielding, A. and Bell, J., 1997, A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, pp. 38–49.

Forester, B.C., 1985, An examination of some problems and solutions in monitoring urban areas from satellite platforms. *International Journal of Remote Sensing*, **6**, pp. 139–151.

Githiru, M. and Lens, L., 2004, Using scientific evidence to guide the conservation of a highly fragmented and threatened Afrotropical forest. *Oryx*, **38**, pp. 404–409.

Guisan, A. and Zimmermann, N.E., 2000, Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, pp. 147–186.

Gustafson, E.J., Hammer, R.B., Radeloff, V.C. and Potts, R.S., 2005, The relationship between environmental amenities and changing human settlement patterns between 1980 and 2000 in the Midwestern USA. *Landscape Ecology*, **20**, pp. 773–789.

Haralick, R.M., Shanmugam, K. and Dinstein, I., 1973, Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, **3**, pp. 610–621.

Hastie, T. and Tibshirani, R., 1990, *Generalized Additive Models* (London, UK: Chapman & Hall).

Holm, M., Lohi, A., Rantasuo, M., Väätäinen, S., Höyhtyä, T., Puumalainen, J., Sarkeala, J. and Sedand, F., 1999, Creation of large mosaics of airborne digital camera imagery. In *Proceedings of the 4th International Airborne Remote Sensing Conference and Exhibition / 21st Canadian Symposium on Remote Sensing*, 21–24 June 1999, Ottawa, Ontario, Canada, Vol. II (Ann Arbor, MI: ERIM International), pp. 520–526.

Hsu, S.Y., 1971, Population estimation. *Photogrammetric Engineering*, **37**, pp. 449–454.

Hutchinson, M.F., 1989, A new procedure for griding elevation and streamline data with automatic removal of spurious pits. *Journal of Hydrology*, **106**, pp. 211–232.

Hutchinson, M.F., 1995, Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Science*, **9**, pp. 305–403.

Jaetzold, R. and Schmidt, H., 1983, *Farm Management Handbook of Kenya II (Part C). Natural Conditions and Farm Management Information, East Kenya* (Nairobi: Ministry of Agriculture).

Kumar, L., Skidmore, A.K. and Knowles, E., 1997, Modelling topographic variation in solar radiation in a GIS environment. *International Journal of Geographical Information Science*, **11**, pp. 475–497.

Landis, J. and Koch, G., 1977, The measurement of observer agreement for categorical data. *Biometrics*, **33**, pp. 159–174.

Lehmann, A., Overton, J.M.C. and Leathwick, J.R., 2002, GRASP: generalized regression analysis and spatial predictions. *Ecological Modelling*, **157**, pp. 189–207.

Li, G. and Weng, Q., 2005, Using Landsat ETM+ imagery to measure population density in Indianapolis, Indiana, USA. *Photogrammetric Engineering & Remote Sensing*, **71**, pp. 947–958.

Lindgren, D.T., 1971, Dwelling unit estimation with color-IR photos. *Photogrammetric Engineering*, **37**, pp. 373–378.

Lo, C.P., 1986, Accuracy of population estimation from medium-scale aerial photography. *Photogrammetric Engineering and Remote Sensing*, **52**, pp. 1859–1869.

Lo, C.P., 1989, A raster approach to population estimation using high-altitude aerial and space photographs. *Remote Sensing of Environment*, **27**, pp. 59–71.

Lo, C.P., 1995, Automated population and dwelling unit estimation from high-resolution satellite images: a GIS approach. *International Journal of Remote Sensing*, **16**, pp. 17–34.

Maggini, R., Lehmann, A., Zimmermann, N.E. and Guisan, A., 2006, Improving generalized regression analysis for the spatial prediction of forest communities. *Journal of Biogeography*, **33**, pp. 1729–1749.

McCullagh, P. and Nelder, J.A., 1989, *Generalized Linear Models* (2nd edn) (New York: Chapman & Hall)

Moore, I.D., Grayson, R.B. and Ladson, A.R., 1991, Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, **5**, pp. 3–30.

Moran, M.S., Bryant, R., Thome, K., Ni, W., Nouvellon, Y., Gonzalez-Dugo, M.P., Qi, J. and Clarke, T.R., 2001, A refined empirical line approach for reflectance retrieval from Landsat-5 TM and Landsat-7 ETM+. *Remote Sensing of Environment*, **78**, pp. 71–82.

Mubareka, S., Ehrlich, D., Bonn, F. and Kayitakire, F., 2008, Settlement location and population density estimation in rugged terrain using information derived from Landsat ETM and SRTM data. *International Journal of Remote Sensing*, **29**, pp. 2339–2357.

Pellikka, P., 1998, Development of correction chain for multispectral airborne video camera data for natural resource assessment. *Fennia*, **176**, pp. 1–110.

Pellikka, P., Lötjönen, M., Siljander, M. and Lens, L., 2009, Airborne remote sensing of spatiotemporal change (1955–2004) in indigenous and exotic forest cover in the Taita Hills, Kenya. *International Journal of Applied Earth Observation and Geoinformation*, **11**, pp. 221–232.

Porter, P.W., 1956, Population distribution and land use in Liberia. PhD thesis, London School of Economics and Political Science, UK.

R Development Core Team, 2008, *R: a Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing). Available online at: http://www.R-project.org.

Republic of Kenya, 2001, *The 1999 Population & Housing Census* (Kenya: Central Bureau of Statistics, Ministry of Planning and National Development).

Ruotsalainen, A., 2008, Enhancing local livelihoods in Taita Hills, Kenya: indigenous tree species as part of farmers´ livelihoods and environmental rehabilitation. MSc thesis, Faculty of Science, Department of Geography, University of Helsinki, Finland.

Schnaiberg, J., Riera, J., Turner, M.G. and Voss, P.R., 2002, Explaining human settlement patterns in a recreational lake district: Vilas County, Wisconsin, USA. *Environmental Management*, **30**, pp. 24–34.

Segurado, P. and Araújo, M., 2004, An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, pp. 1555–1569.

Shaban, M.A. and Dikshit, O., 2001, Improvement of classification in urban areas by the use of textural features: the case study of Lucknow city, Uttar Pradesh. *International Journal of Remote Sensing*, **22**, pp. 565–593.

Soini, E., 2005, Livelihood capital, strategies and outcomes in the Taita Hills of Kenya. *ICRAF Working Paper no. 8* (Nairobi, Kenya: World Agroforestry Centre).

St-Louis, V., Pidgeon, A.M., Radeloff, V.C., Hawbaker, T.J. and Clayton, M.K., 2006, High-resolution image texture as a predictor of bird species richness. *Remote Sensing of Environment*, **105**, pp. 299–312.

Swets, K., 1988, Measuring the accuracy of diagnostic systems. *Science*, **240**, pp. 1285–1293.

Teillet, P.M., Guindon, B. and Goodenough, D.G., 1982, On the slope-aspect correction of multispectral scanner data. *Canadian Journal of Remote Sensing*, **8**, pp. 84–106.

Thuiller, W., Araújo, M.B. and Lavorel, S., 2004, Do we need land-cover data to model species distributions in Europe? *Journal of Biogeography*, **31**, pp. 353–361.

Tobler, W., Deichmann, U., Gottsgen, J. and Maloy, K., 1995, *The Global Demography Project*. Technical Report TR-6-95 (Santa Barbara, CA: National Center for Geographic Information and Analysis (NCGIA), University of California).

Tobler, W., Deichmann, U., Gottsegen, J. and Maloy, K., 1997, World population in a grid of spherical quadrilaterals. *International Journal of Population Geography*, **3**, pp. 203–225.

United Nations Environment Programme (UNEP) / Global Resource Information Database (GRID), 2006, UNEP/GRID spatial data clearinghouse. Available online at: http://na.unep.net/siouxfalls/datasets/datalist.php (accessed 19 November 2009).

Venables, W.N. and Ripley, B.D., 2002, *Modern Applied Statistics with S* (4th edn) (New York: Springer-Verlag)

Weisberg, S., 1980, *Applied Linear Regression* (New York: Wiley).

Zaniewski, A.E., Lehmann, A. and Overton, J.M., 2002, Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, pp. 261–280.

Zar, J.H., 1999, *Biostatistical Analysis* (4th edn) (London, UK: Prentice-Hall).

Zimmermann, N.E., 2000, Shortwavc.aml. Available online at: http://www.wsl.ch/staff/niklaus.zimmermann/programs/aml1_1.html (accessed 01 January 2009).

Zimmermann, N.E., Edwards JR., T.C., Moisen, G.G., Frescino, T.S. and Blackard, J.A., 2007, Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species in Utah. *Journal of Applied Ecology*, **44**, pp. 1057–1067.