Forest Research

Area estimation by double-calibration of a remote-sensed feature map to fieldwork observations via manual image interpretation

By Alan Brewer
National Forest Inventory
Statistician

The situation

- We have a digital spatial dataset (digital map) representing features of interest, covering the whole country, that has been generated from automated interpretation of remote sensing data
- This dataset is not entirely accurate, with a proportion of mis-identified features, missed features, and locational errors of successfully identified features
- In addition to this generated spatial dataset, aerial photography is available with full coverage across the country
- We wish to conduct spatial sample surveys (using GIS-based software) to ground-truth and assess the quality of this dataset
- We also wish to use the results of these surveys to relate to the digital map in order to obtain estimates of known accuracy of total areas of classes of features in geographical sub-populations

The approach

- A statistically-designed sample survey of small areas is selected across the population within which the aerial photography is visually assessed by photographic interpreters and relevant features are digitally mapped.
- The results of this desk exercise are referred to as hand-mapped data.
- In addition, for a subset of these samples, surveyors are sent to the sample sites to ground-truth and correct errors in the hand-mapped data.
- There are therefore 3 spatial datasets which each have information to use for comparison and estimation:
  1. The digital map covering the whole country
  2. Spatial hand-mapped data within a sample of small areas
  3. Spatial data from field surveys on a sub-sample of the hand-mapped sample

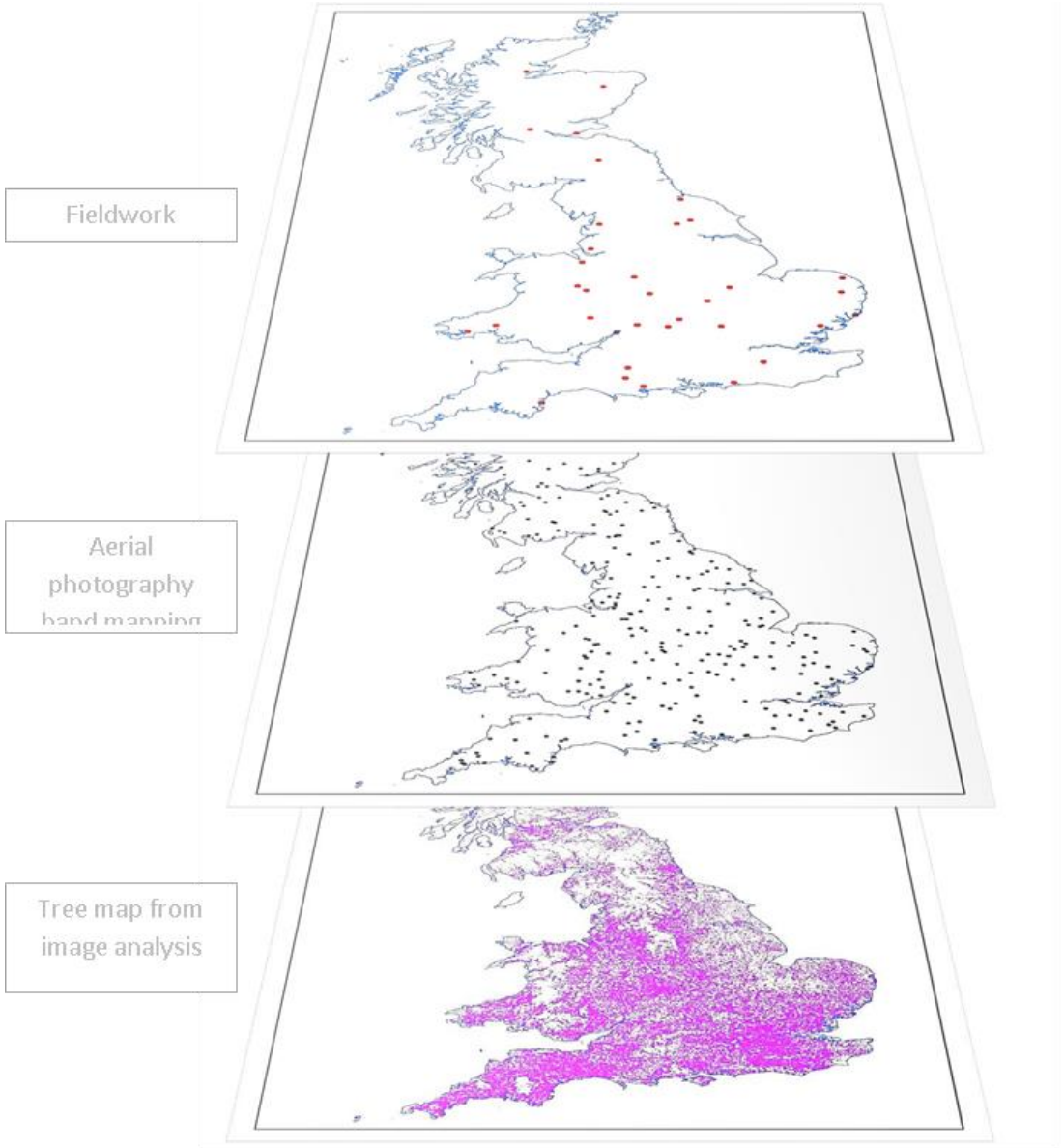
The problem

- With 3 separate spatial datasets, how can we combine the information in each to generate efficient statistical estimates of total areas of classes of features?
- Also, can we quantify the accuracy of the resulting estimates?

The solution

- The first stage is to use GIS software to superimpose 2 spatial datasets to quantify their intersecting and non-intersecting areas.
- This is performed twice, to compare the digital map to the hand-mapped data, and secondly to compare the hand-mapped data to the fieldwork results
- Using the results of these GIS analyses, a statistical technique has been devised to double-calibrate from digital map to hand-mapping and from hand-mapping to fieldwork to produce fieldwork-based area estimates and standard errors

Combining fieldwork, hand-mapping samples and the NTM map

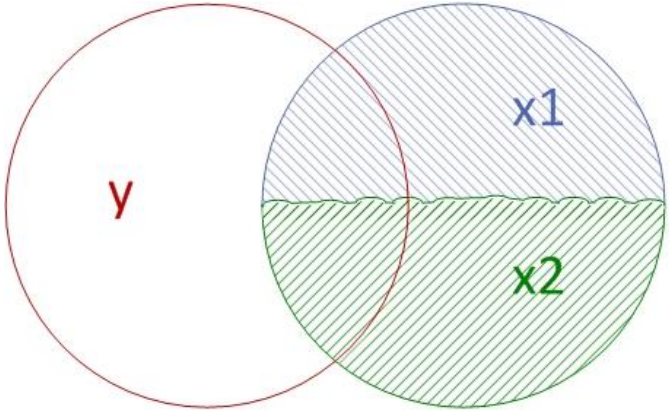


GIS analysis

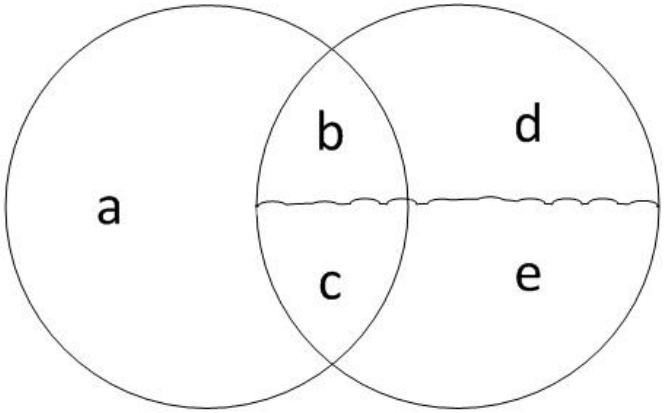
- Each spatial dataset contains a number of classes of features, with areas of individual features represented by polygons
- The spatial dataset to be calibrated is referred to here as the 'primary dataset' that is to be calibrated by the 'calibrating dataset'
- The sets of classes of features represented in each of the datasets may or may not be the same in both datasets
- The GIS analysis involves separately superimposing the spatial representations in the primary dataset onto each of the classes of features in the calibrating dataset
- By this means, each class of features in the calibrating dataset is separately related to the classes of features in the primary dataset

A schematic example in which the primary dataset has 2 classes of features

Sample

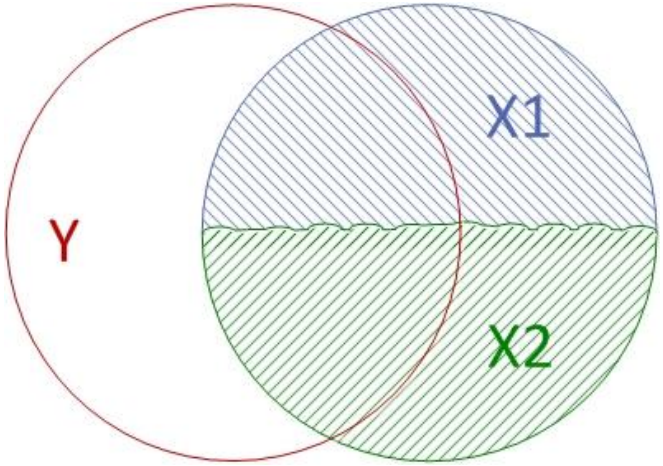


(a) Sample-level schematic relationship between primary and calibrating spatial datasets



(b) Decomposition of (a) into areas of intersection and non-intersection

Population



Statistical analysis

- With a primary dataset with k classes, within a single sample i the total area of polygons of a particular class of features j in the calibrating dataset, $y_{i,j}$ is made up of:

$$y_{i,j} = a_{i,j} + \sum_{k} c_{i,j,k}$$

where $a_{i,j}$ is the area of the polygons not intersecting with the primary dataset and the $c_{i,j,k}$ are the areas of intersection with each of the k classes of the primary dataset.

- These, together with the total areas of polygons $x_{i,k}$ of each of the k classes of the primary dataset within the sample comprise the data to be statistically analysed.
- From these, we can sum across the sample as a whole to obtain:

$$\boldsymbol{y_j} = \sum s_{i,j} \quad \boldsymbol{a_j} = \sum a_{i,j} \quad \boldsymbol{c_{j,k}} = \sum c_{i,j,k} \quad \boldsymbol{x_k} = \sum x_{i,,k}$$

- These summations across the sample are then used to parameterise the calibrating equations:

Statistical analysis

From the totals across the sample, we derive 'scaling-up' and 'scaling-down' calibrating ratios:

$s_j = y_j/(y_j - a_j)$ (the 'scaling-up parameter) and;

$r_{j,k} = c_{j,k}/x_k$ (the 'scaling down' parameters for each class of features in the primary dataset)

We can then use these to define the statistical model at sample site level:

$$y_{i,j} = s_j * \sum_k (r_{j,k} * x_{i,k}) + \varepsilon_{i,j}$$

where:

$y_{i,j}$ is the area of calibrating dataset features of category i in sample square j

$x_{j,k}$ is the area of primary dataset features of category k in sample square j

$\varepsilon_{i,j}$ are random errors associated with areas of calibrating dataset category i in sample square j

$r_{i,k}$ and s_i and are the calibration parameters calculated across the sample as a whole.

Statistical analysis

- We have estimates of total areas X_k in the population of each of the classes of features in the primary dataset and we can use these and the values of the 'scaling-up' and 'scaling-down' calibration parameters to estimate population values of each of the classes of features in the calibrating dataset:

$$Y'_i = s_i * \sum_k (r_{i,k} * X_k)$$

- The variances of these estimators are obtained by considering the statistical model of errors at the sample site level and use the formula for the asymptotic variance of a function of random variables:

  $Var(Y') = \mathbf{f'Vf}$

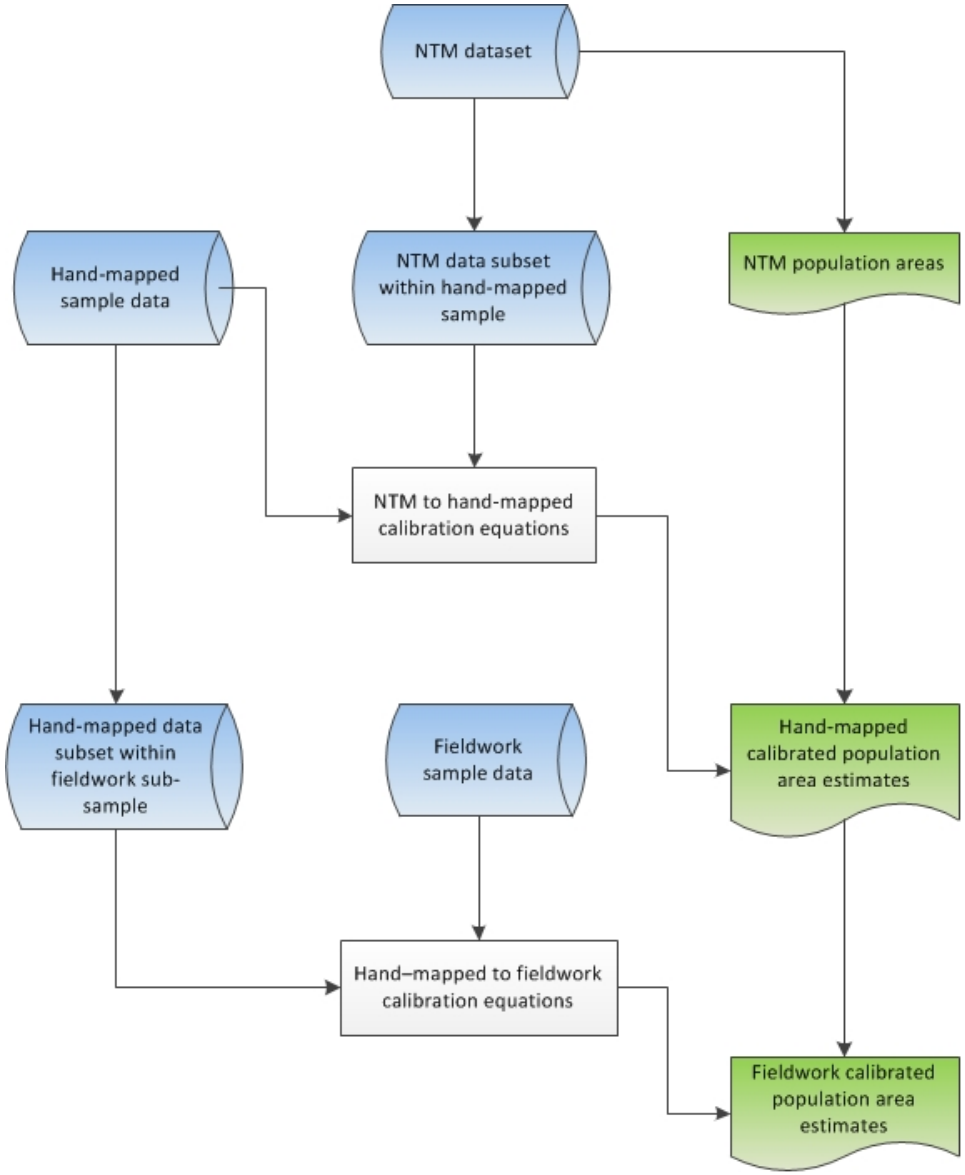
  where **V** is the variance-covariance matrix of the basic variables generated by the GIS analysis and **f** is the vector of first derivatives of the function of these variables involved in the above population estimation equation.

Double calibration

The GIS and statistical analysis calibration processes are applied twice (as described earlier).

This results in the following schema for the overall analysis:

The double-calibration process



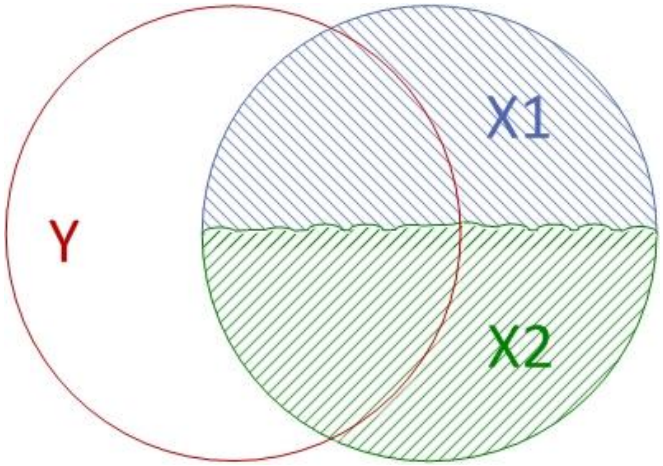
Considerations

In applying this approach to estimate total population areas of spatial features, there are three considerations that we may want to investigate:

1. Does the application of this spatial approach in a given situation result in better estimates than a non-spatial approach? (e.g. straight regression or ratio estimators)
2. Do we gain accuracy by effectively stratifying the population and applying the approach separately within each stratification level?
3. Considering relative costs of collecting hand-mapped and fieldwork samples, is there an optimum ratio in the sizes of these samples?

Spatial versus non-spatial analysis

The value of using spatial information in this way is basically determined by the degree of coincidence between the spatial datasets. Considering the population representation of the spatial relationship between the datasets involved in the analysis:



Considering the extremes:

- If there is no area of intersection the spatial approach will yield no information about the calibrating dataset
- If the datasets are exactly coincident then the analysis is not needed since there is exact identity
- So in general the power of the approach increases with the amount of coincidence between the datasets used in the analysis

Stratification

- As in non-spatial sampling, if the overall population can be sub-divided (geographically) where it is expected that different classes of the stratification factor(s) contain different spatial relationships between the datasets, separate calibrations within these stratification classes can greatly improve overall performance
- Any benefits can be observed by comparing the variances or se's from analyses that take account of the identified stratification against those obtained without using the stratification.
- Different stratifications can be used in each of the two calibration analyses in the double-calibration process

Optimum relative sample sizes

- By assuming independence between the two stages of the double calibration process, the overall variances of the final fieldwork-based estimators are composed of the sum of 2 variances deriving from each stage of the double calibration
- (However, the assumption of independence is an approximation since the hand-mapped data in the fieldwork sub-sample is used in both processes, so there is actually some amount of association between the individual calibration analyses)
- With information on the costs of collecting hand-mapped and fieldwork samples, the solution is to minimise:

$$v(m_1, m_2) = \frac{v_1}{m_1} + \frac{v_2}{m_2}$$

where v_1 and v_2 are the resulting *per sample* variances of the 2 calibration processes and m_1 and m_2 are the respective sample sizes, subject to:

$$c_1 m_1 + c_2 m_2 = K$$

Using Lagrangian multipliers, the solution to this is:

$$\frac{m_1}{m_2} = \sqrt{(v_2 c_1 / v_1 c_2)}$$

Case study

Areas of small woodlands and trees outside of NFI woodland

Objectives

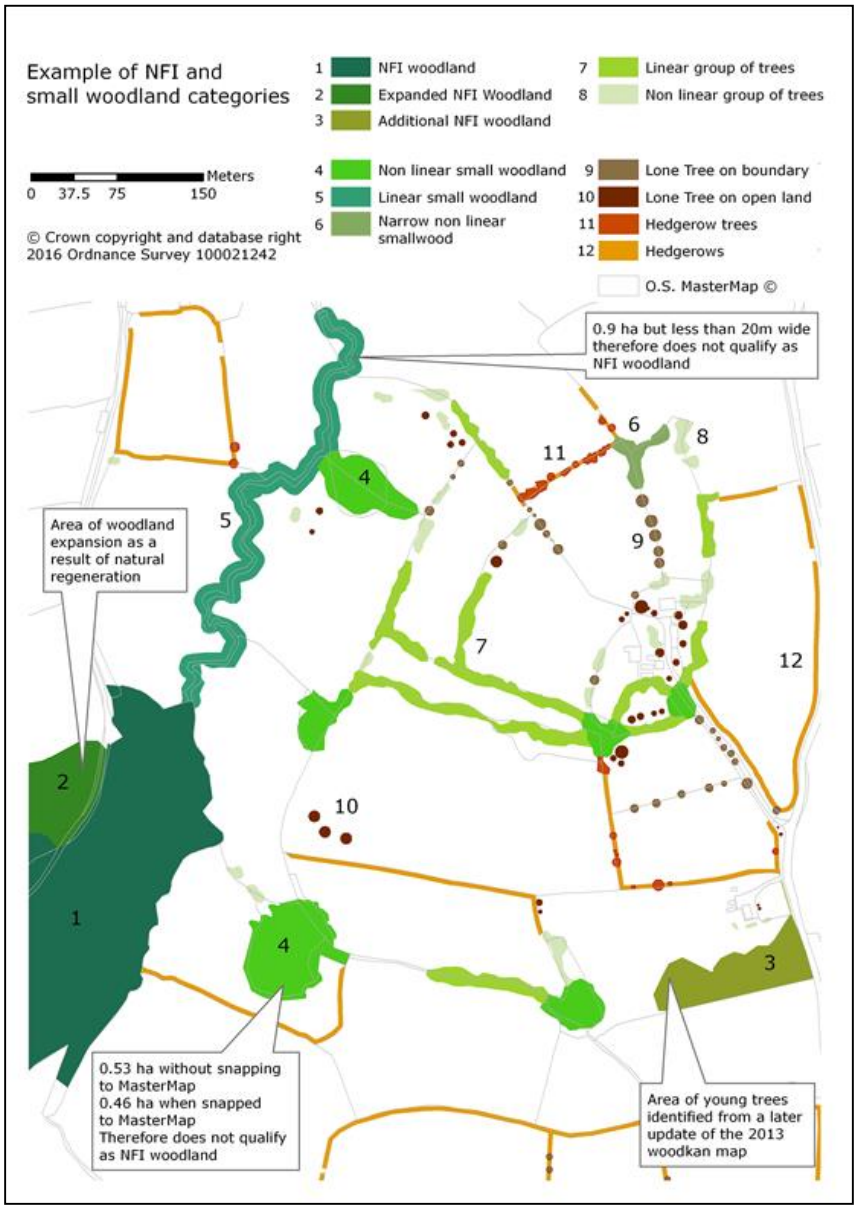
To estimate areas of features of small woodlands, groups of trees and individual trees outside of main woodland areas in England and Wales*

* >0.5 hectares and >20 metres wide

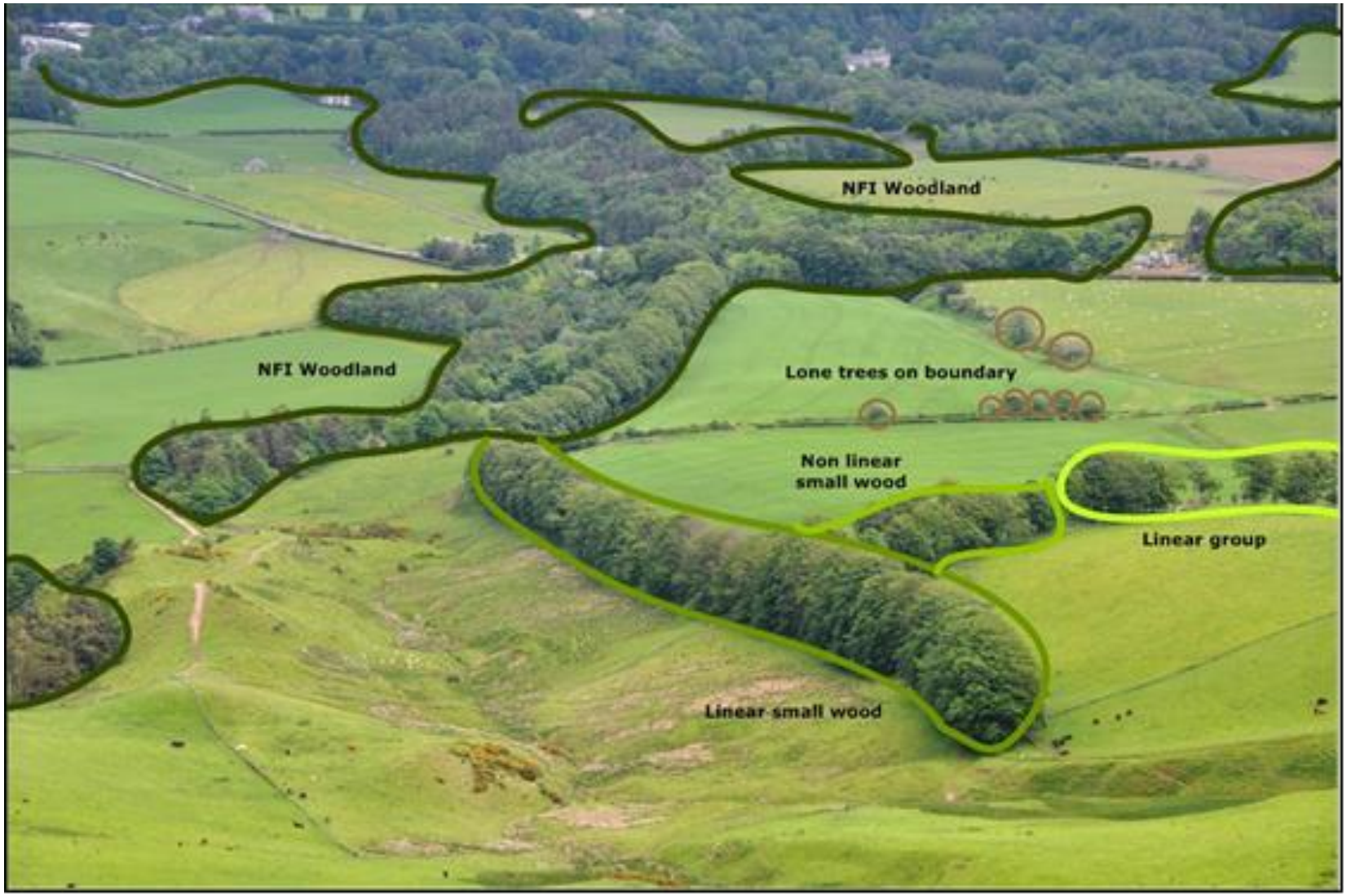
Data sources

1. National Tree Map produced by Blue Sky International:
   - Automated map of tree canopy cover generated using bespoke image-processing techniques of spatial datasets
2. Sample of 1 km. squares within which tree cover features were assessed using visual interpretation of aerial photography
3. Sub-sample of the 1 km. squares visited by field surveyors for ground-truthing of the aerial photography interpretation work

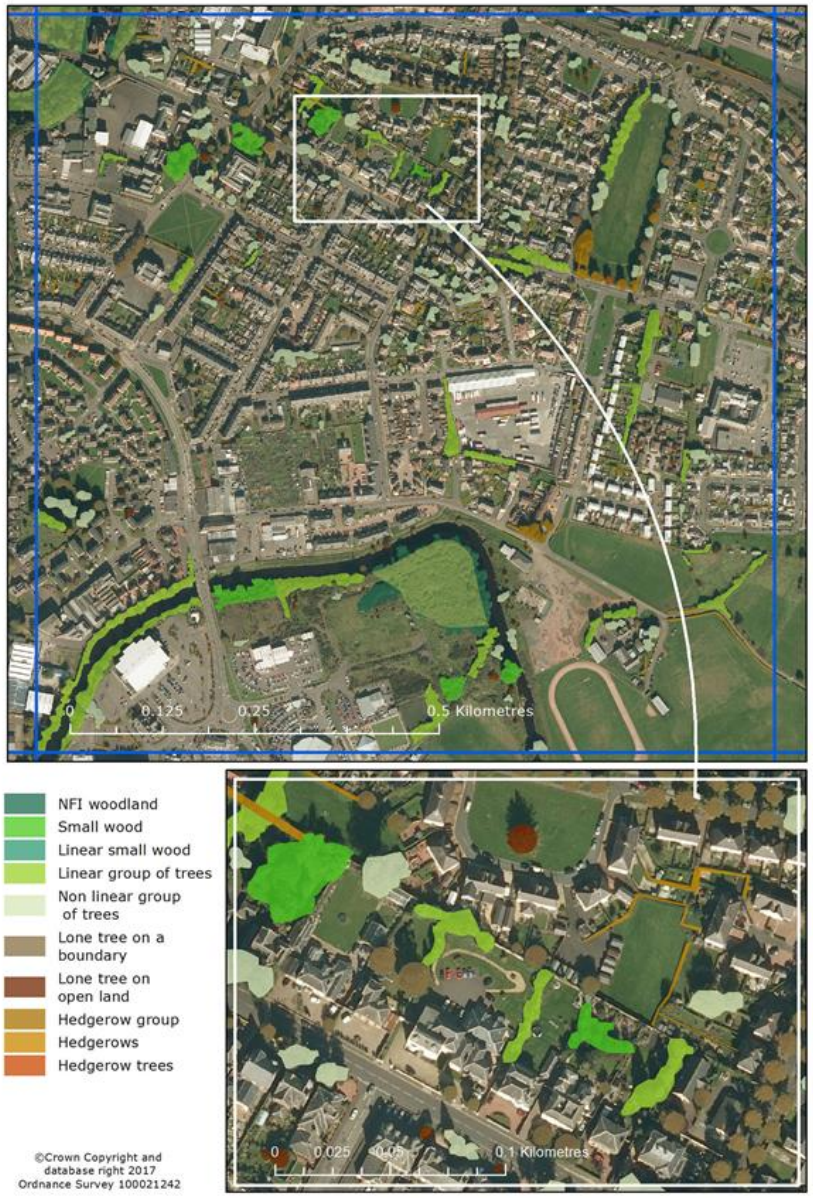
Schematic map of classes of small wood and tree features



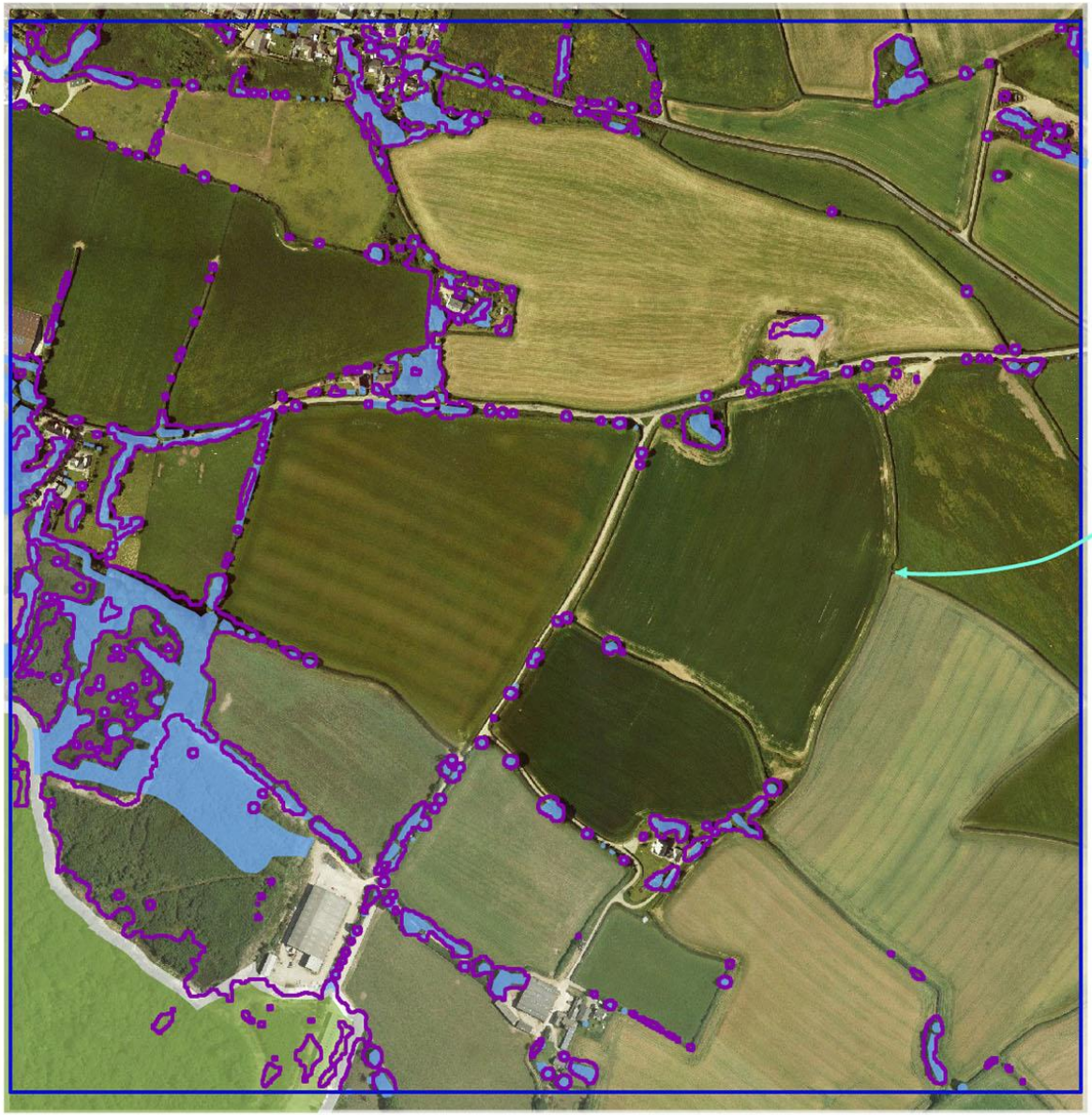
Examples of woodland and non-woodland features



Section of a fieldwork sample square with marked features



A 1 km. square hand-mapped sample



Hand-mapped tree cover polygons
Automated small woods polygons
1km-by-1km callibration square
National Forest Inventory woodland ≥0.5ha

0 0.125 0.25 Kilometres

Sample sizes and stratification

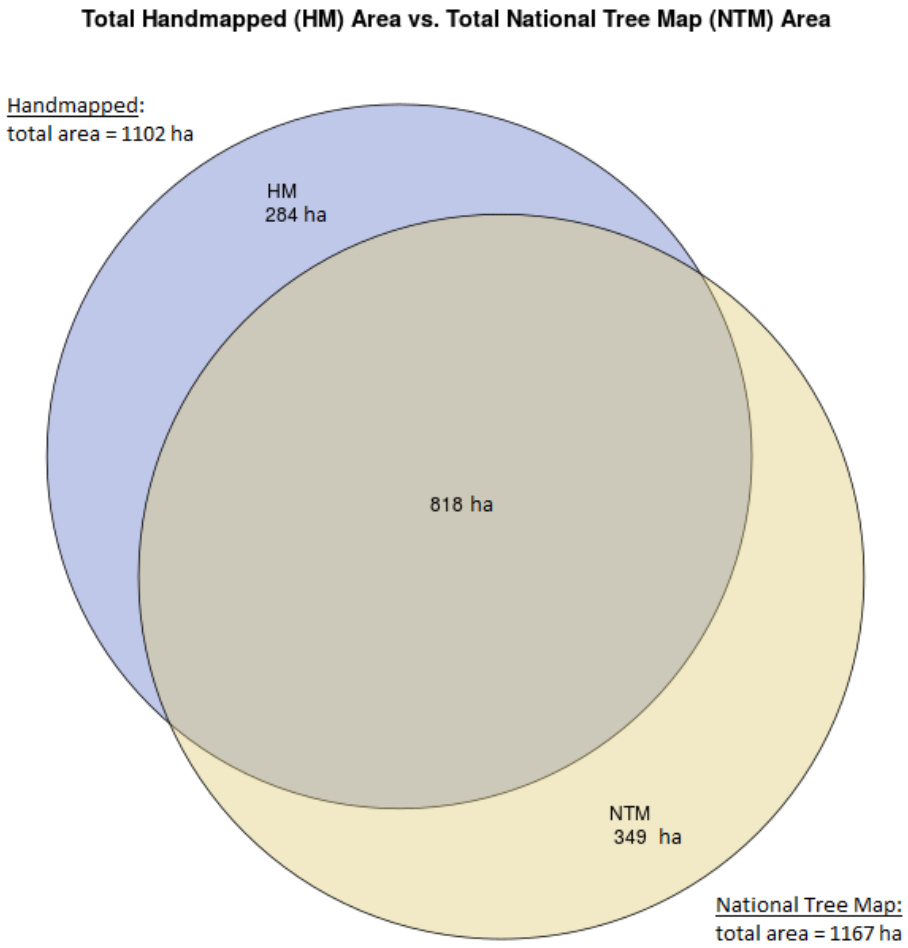
Country/Region	Hand-mapped sample			Fieldwork sub-sample		
	Sample squares containing:		Total sample	Sample squares containing:		Total sample
	Rural areas	Urban areas	squares	Rural areas	Urban areas	squares
GB total	**271**	**101**	**277**	**34**	**16**	**36**
England and Wales	**212**	**87**	**217**	**29**	**14**	**31**
England	**185**	**78**	**190**	**25**	**13**	**27**
North West England	22	15	23	2	2	2
North East England	13	6	13	1	1	1
Yorkshire and the Humber	22	6	22	2	1	2
East Midlands	24	8	24	2	2	2
East England	25	11	26	4	1	4
South East and London	28	15	31	5	1	5
South West England	33	11	33	5	3	6
West Midlands	18	6	18	4	2	5
Wales	**27**	**9**	**27**	**4**	**1**	**4**
Scotland	**59**	**14**	**60**	**5**	**2**	**5**
North Scotland	6	1	6	1	0	1
North East Scotland	9	2	10	1	1	1
East Scotland	10	3	10	1	0	1
South Scotland	24	7	24	1	1	1
West Scotland	10	1	10	1	0	1

- In the analyses, the NTM to hand-mapping calibration was stratified by urban and rural and by regions in England + Wales
- The Scotland data was not used in the NTM to hand-mapping calibration
- The hand-mapping to fieldwork calibration was stratified only by urban and rural categories
- All GB data was used for the hand-mapping to fieldwork calibration

Areas of major categories of tree cover outside woodland

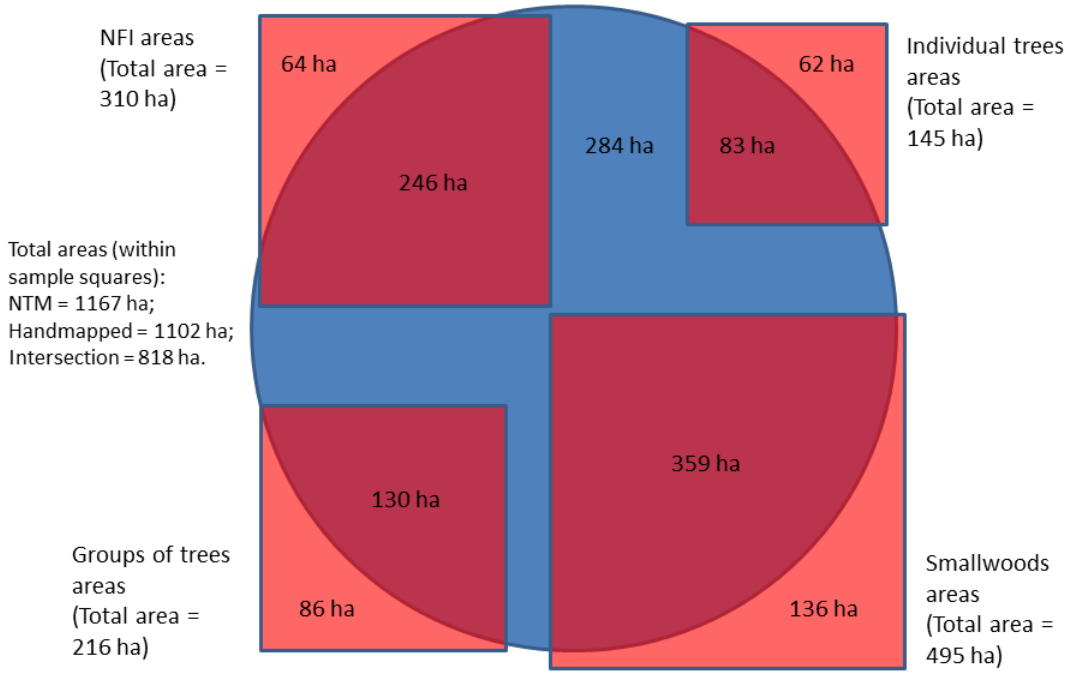
Land category	Total NFI woodland	Small woods		Groups of trees		Lone trees		Total area of woodland and tree cover	
	(000 ha)	(000 ha)	SE%	(000 ha)	SE%	(000 ha)	SE%	(000 ha)	SE%
Great Britain	**3,075**	**390**	***7***	**255**	***6***	**97**	***6***	**3,817**	***5***
Rural	2,984	316	*8*	165	*6*	64	*6*	3,530	*6*
Urban	90	74	*12*	90	*13*	33	*15*	286	*10*
England	**1,336**	**295**	***7***	**193**	***6***	**78**	***7***	**1,901**	***5***
Rural	1,271	238	*8*	125	*9*	52	*6*	1,686	*5*
Urban	65	57	*12*	67	*13*	26	*15*	215	*11*
Scotland	**1,429**	**46**	***21***	**29**	***12***	**9**	***15***	**1,513**	***13***
Rural	1,413	41	*24*	23	*14*	7	*17*	1,484	*15*
Urban	16	5	*24*	7	*26*	2	*31*	30	*18*
Wales	**309**	**49**	***8***	**33**	***9***	**10**	***17***	**402**	***7***
Rural	300	38	*9*	17	*8*	5	*11*	360	*7*
Urban	9	12	*20*	16	*17*	5	*32*	42	*14*

Coincidence between the hand-mapped sample and NTM areas

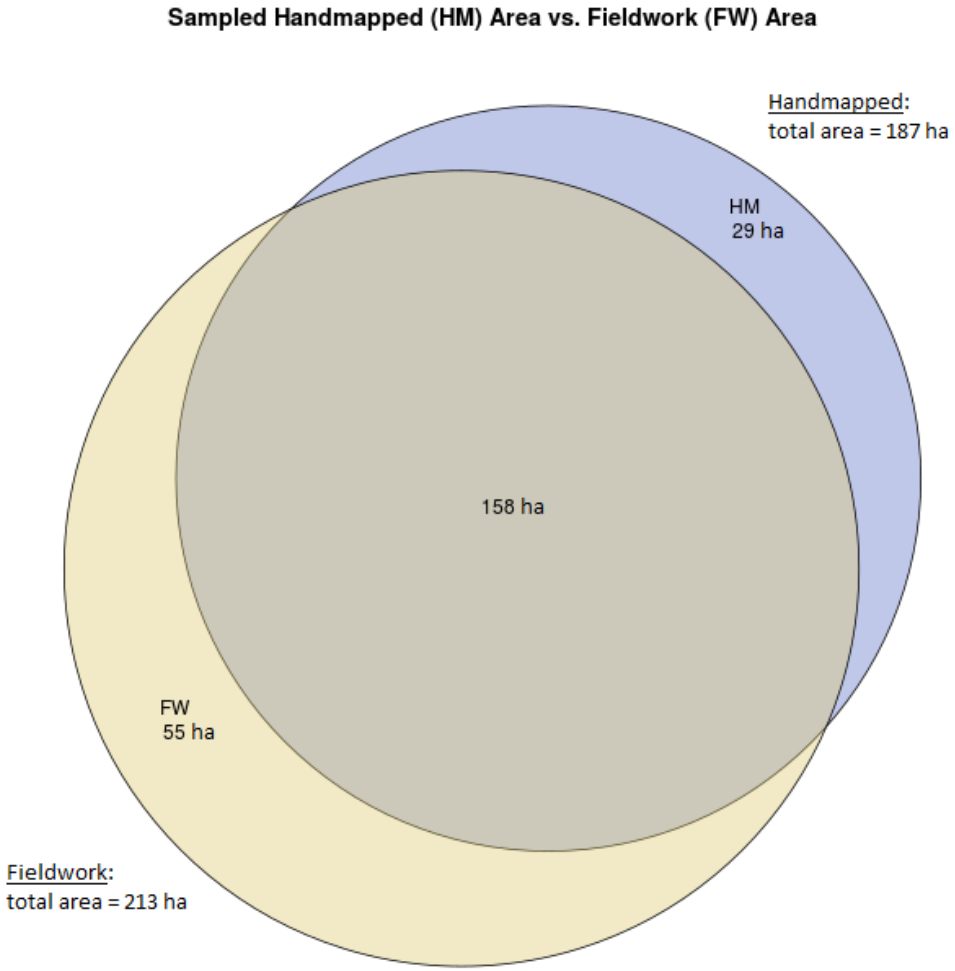


Hand-mapped vs NTM broken down by NTM classes

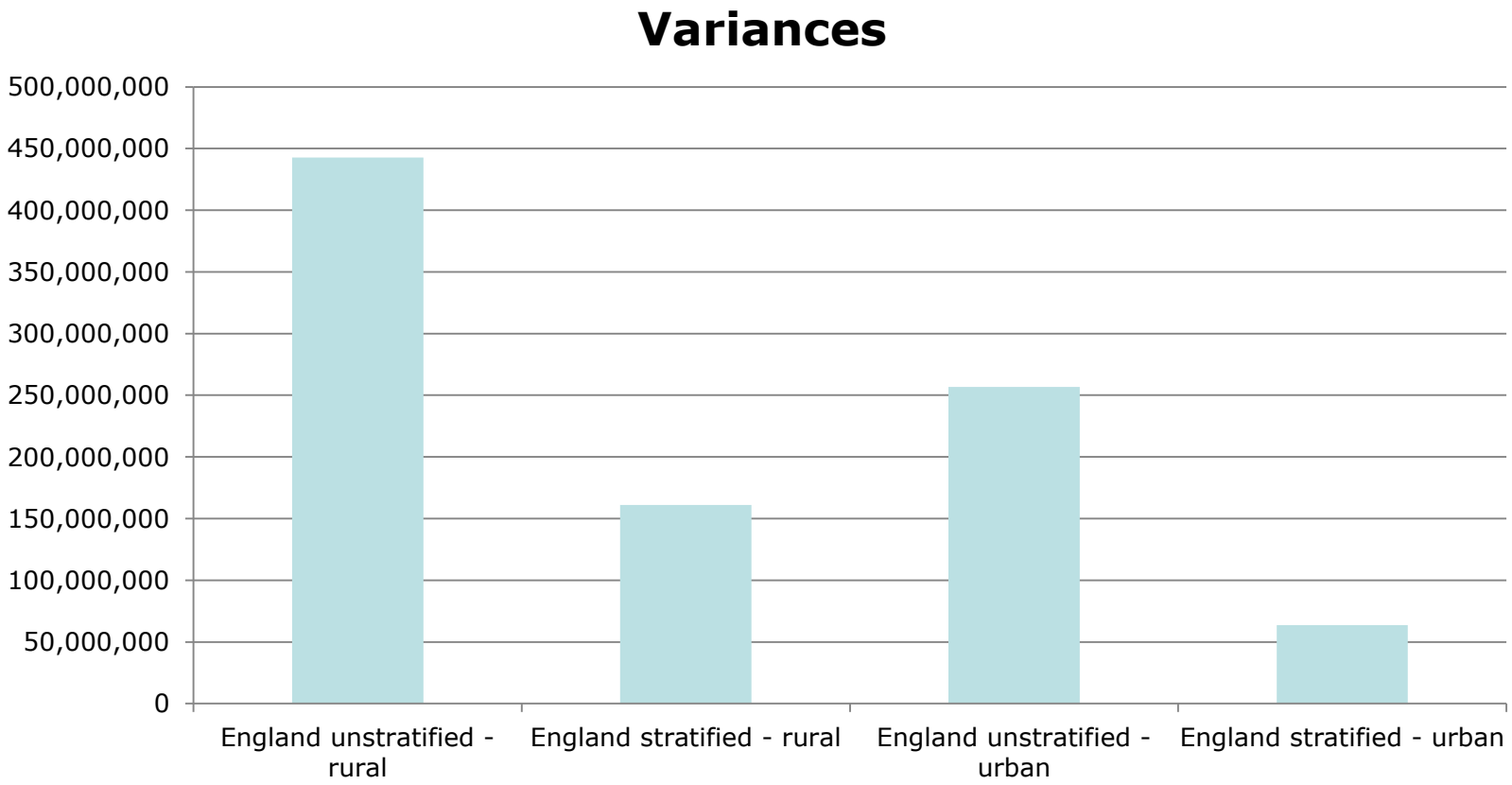
Sampled Handmapped Area (circle) with National Tree Map Areas (squares) by NTM class



Coincidence between handmapped and fieldwork samples



Effect of regional stratification in England on variance of country estimates of total areas of small woodlands and trees



Rural variance ratio: 2.75
Urban variance ratio: 4.00

Sampling efficiency

England and Wales	Rural		Urban	
	Fieldwork	**Handmapped**	**Fieldwork**	**Handmapped**
Cost of sample collection	£500	£140	£500	£140
Variance *per sample* contributed by calibration (v_1 and v_2)	2,558,552,383	21,699,192,365	732,839,454	3,317,952,066
Optimum sample size ratio (handmapped/fieldwork)	**5.5**		**4.0**	
Actual sample numbers	212	34	87	16
Actual sample ratios	**6.2**		**5.4**	

NFI small woodlands and trees report

https://www.forestry.gov.uk/fr/beeh-a2uegs