— Selecting weights of satellite image and ancillary information in k-NN estimation: a genetic algorithm approach —

Erkki Tomppo
Finnish Forest Research Institute
National Forest Inventory
email: erkki.tomppo@metla.fi
www: http://www.metla.fi/projects/vmi/e-tomppo.htm

Merja Halme
Helsinki Scool of Economics
email: merja.halme@hkkk.fi

The contribution of Jouni Peräsaari
Finnish Forest Research Institute
National Forest Inventory
is greatly acknowledged

— FOREST INVENTORY GOAL —

produce statistics and time series about forests concerning

$\Longrightarrow$ Land use and ownership status

$\Longrightarrow$ Sites and their quality

$\Longrightarrow$ Volume, quality and structure of growing stock

$\Longrightarrow$ Increment of growing stock and its variation

- Silvicultural status of forests
- Applied and needed silvicultural and cutting regimes
- Forest health condition
- Biological diversity of forests

$\Longrightarrow$ Typically 100 - 400 variables are measured in the field

— GOAL (cont.) —

for different area units

$\Longrightarrow$ Usually with field plot data

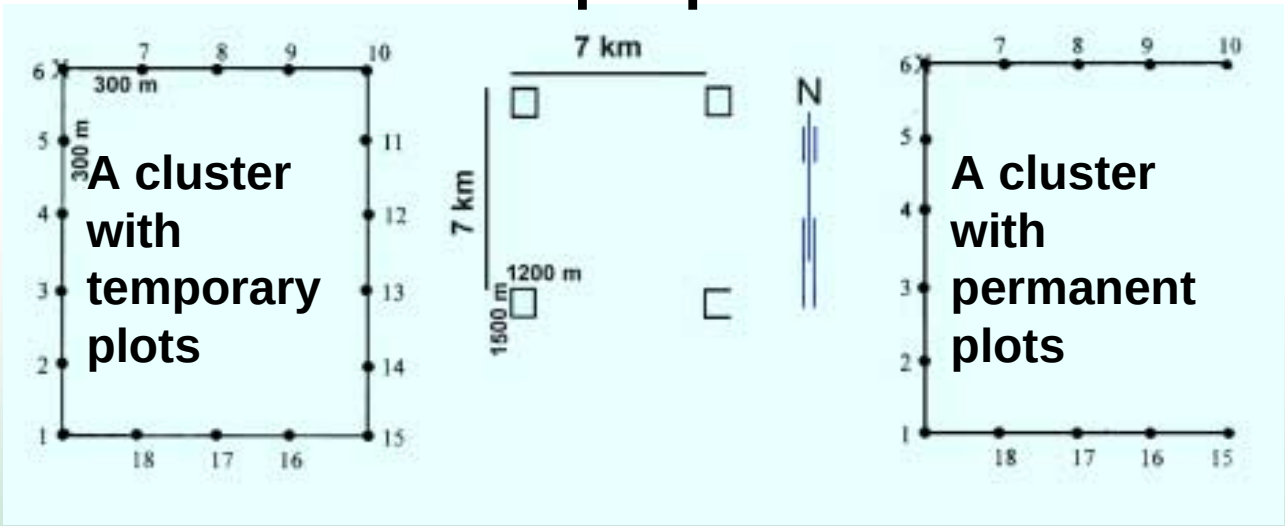
- Whole country ($c \times$ 10 mill. ha)
- Forestry centre ($c \times$ 1 mill. ha)

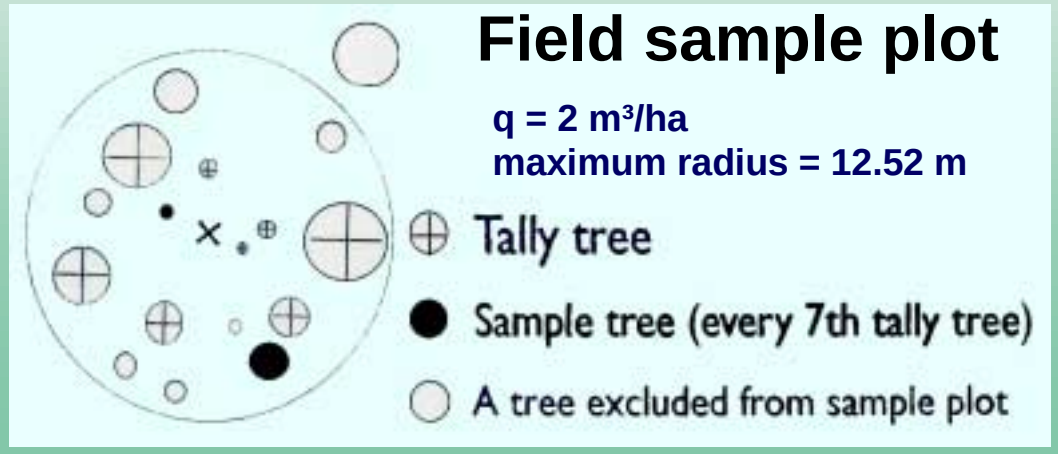
$\Longrightarrow$ With multisource data

- Municipality ($c \times$ 10 000 ha)
- Village ($c \times$ 1000 ha)
- Forest holding ($\sim$ 100 ha)
- In the future, for forest stand ($\sim$ 1 ha) , a single tree

Sampling design of FNFI

Sample plot





— Area and volume estimates, field data —

- Area estimates: land use classes, tree species dominance, quality of forests, age distributions, etc.
- Volume estimates: volumes by timber assortment classes and by strata, increments by tree species, etc.

The goal is to estimate

$$M(A) = \frac{\mathbf{X}}{\mathbf{Y}} = \frac{\int_A x(t)\,dt}{\int_A y(t)\,dt},$$

where

$x(t), t \in R^2$ is the variable of interest, e.g. the indicator of a land use class, the volume of a timber assortment
$y(t), t \in R^2$ is the indicator of a stratum (e.g. forestry land) and
$A \subset R^2$ a computational unit (e.g. forestry centre)

— Area and volume estimates, field data —

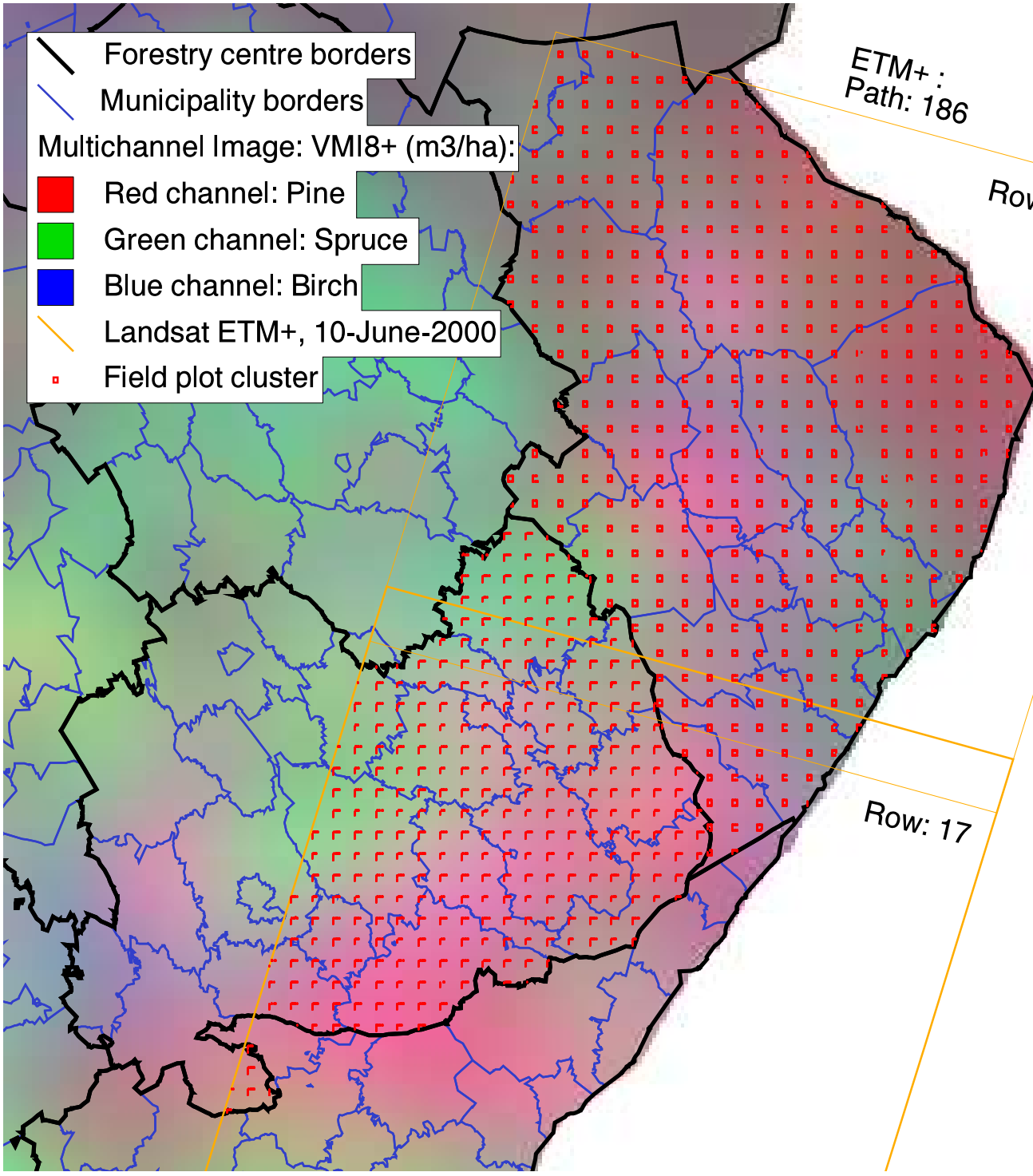
x_i ja y_i be the observed values on the plot i.
The ratio estimator of $\mathbf{M}$:

$$m_n = \frac{\sum_i^n x_i}{\sum_i^n y_i} = \frac{\bar{x}}{\bar{y}},$$

n is the number of field plots in the area A.
The estimator for the totals is $m_n \times |A|$.

- How to assess the reliability ?
- The error variance $E(m_n - M(A))^2$.
- Unbiased estimator not known.
  - spatial correlation
  - systematic sampling
- Biased, conservative estimators can be derived using the properties of second order stationary stochastic processes (Matérn 1960).

Ancillary information, e.g., satellite image information makes it possible to compute estimates for areas of, typically, 10 000 ha instead of 200 000 - 1 mill. ha with field data only.



— A non-paramteric k-nn estimation with covariate, e.g. satellite images —

$\Longrightarrow$ distance measure d in the *covariate space* (e.g., the Euclidean distance in R^6, when using the 6 spectral channels of a Landsat TM image).

$\Longrightarrow$ for the pixel p to be analysed, compute $d_{p_i,p} = \|p_i - p\|$ to each pixel p_i whose ground truth is known (to pixel with field plot i).

$\Longrightarrow$ for $k \approx$ 5-10, let $p_{(1)}, p_{(2)}, \ldots, p_{(k)}$ be the k nearest field plot pixels with respect to d.

$\Longrightarrow$ The weight of field plot i to pixel p is defined as

$$w_{i,p} = \frac{1}{d^2_{p_i,p}} \Big/ \sum_{j=1}^{k} \frac{1}{d^2_{p_{(j)},p}}$$

if pixel p_i is among the k nearest to p, otherwise $w_{i,p} = 0$.

— Estimation —

In the image analysis process, weights $w_{i,p}$ are summed over pixels p by computation units. The weight of the field plot i to computation unit u is defined as

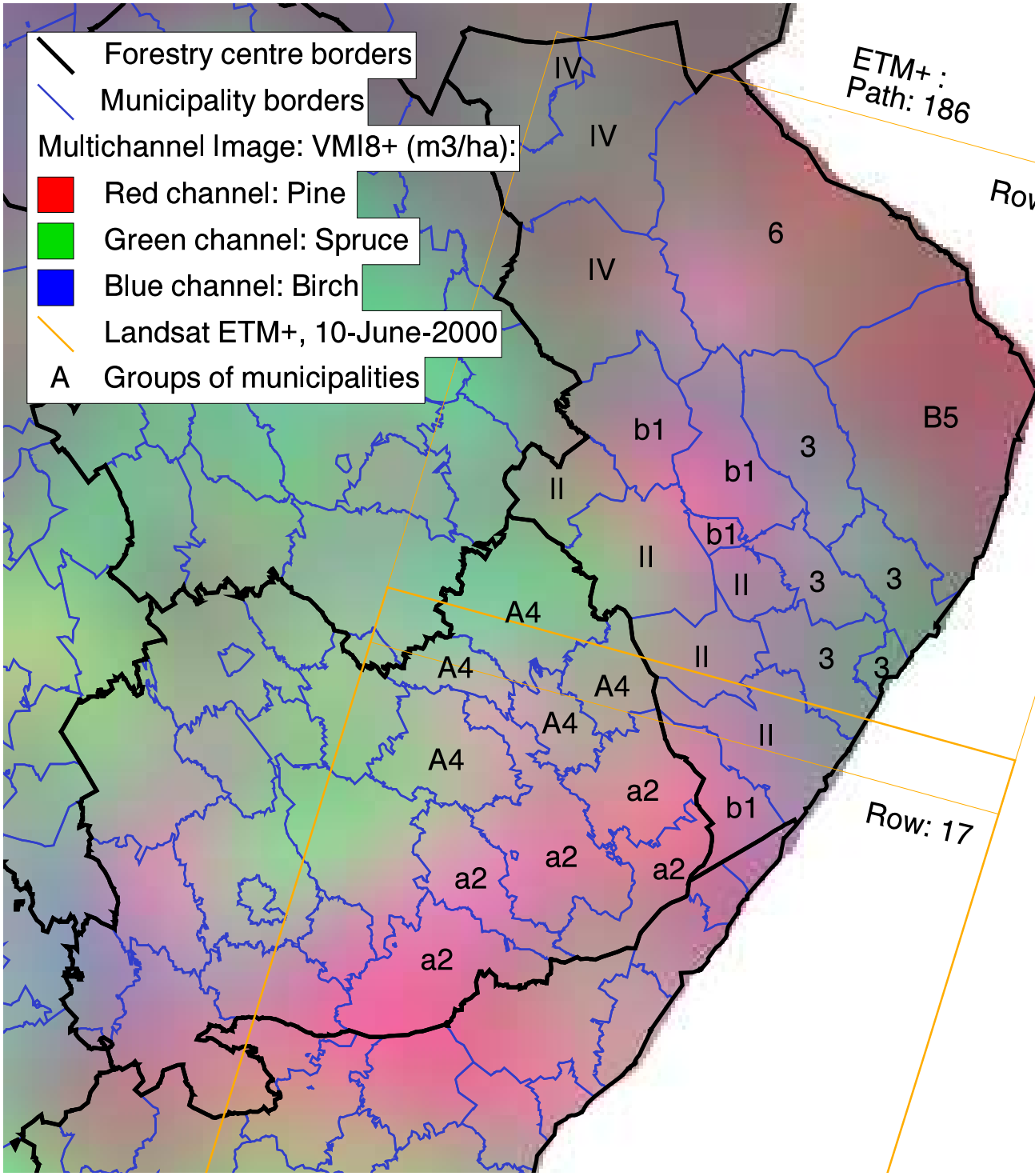
$$c_{i,u} = \sum_{p \in u} w_{i,p}$$

The estimation returns to the estimation with pure field data, e.g. mean timber volumes estimates by forestry land strata are

$$v = \frac{\sum_{i \in I_s} c_{i,u} v_{i,t}}{\sum_{i \in I_s} c_{i,u}},$$

where $v_{i,t}$ is the volume per hectares of the timber assortment t in sample plot i.

Large scale variation of forest variables and the municipality groups



— Problems —

1. How to select the geographical area from which the nearest neighbours are sought

   - old method: moving window for avoiding neighbours from different vegetation zone
   - new method: large scale forest variables

2. What kind of features (image variables) should be used in the estimation, original bands, transformations
3. What is the distance function
   - Eudlidean distance, in the product space of image variables and large scale forest variables
   - the weights of elements by means of a genetic algorithm

— Distance function —

$$d^2_{p_i,p} = \sum_{i=1}^{n_s} \omega_{i,s}(s_{i,p_j} - s_{i,p})^2 + \sum_{i=1}^{n_t} \omega_{i,t}(t_{i,p_j} - t_{i,p})^2$$

where

$s_{i,p}$ are the image variables, in our case, spectral bands and all band ratios
$t_{i,p}$ are the large area forest variables and
n_s and n_t are the numbers of the spectral and large area forest variables
$\omega_{i,s}$ and $\omega_{i,t}$ are the variable weights

The problem:
how to select the weights ?

GENETIC ALGORITHMS

- metaheuristic strategies that direct and modify search heuristics to overcome local optima
- often efficient when no exact methods easily available
- often produce good solutions
- always a great deal of fine tuning needed

Genetic algorithm imitates the evolution.

ALGORITHM IN NUTSHELL

$\Longrightarrow$ random initial population

$\Longrightarrow$ selection (tournament)

$\Longrightarrow$ population (set of parents)

$\Longrightarrow$ crossover producing children - elitism - mutation (guaranteeing diversity)

$\Longrightarrow$ result: new population (= next generation)

$\Longrightarrow$ go to selection

Key parameters

- npop - number of weights vectors in one population
  $\sim$ 50
- cross - probability for elements to be switched in two weights vectors
  $\sim$ 0.75
- muta - mutation probability
  $\sim$ 0.05
- calm - probability of accepting an inferior solution created by mutation
  $\sim$ 0.5
- ngen - number of generations
  $\sim$ 30 - 80

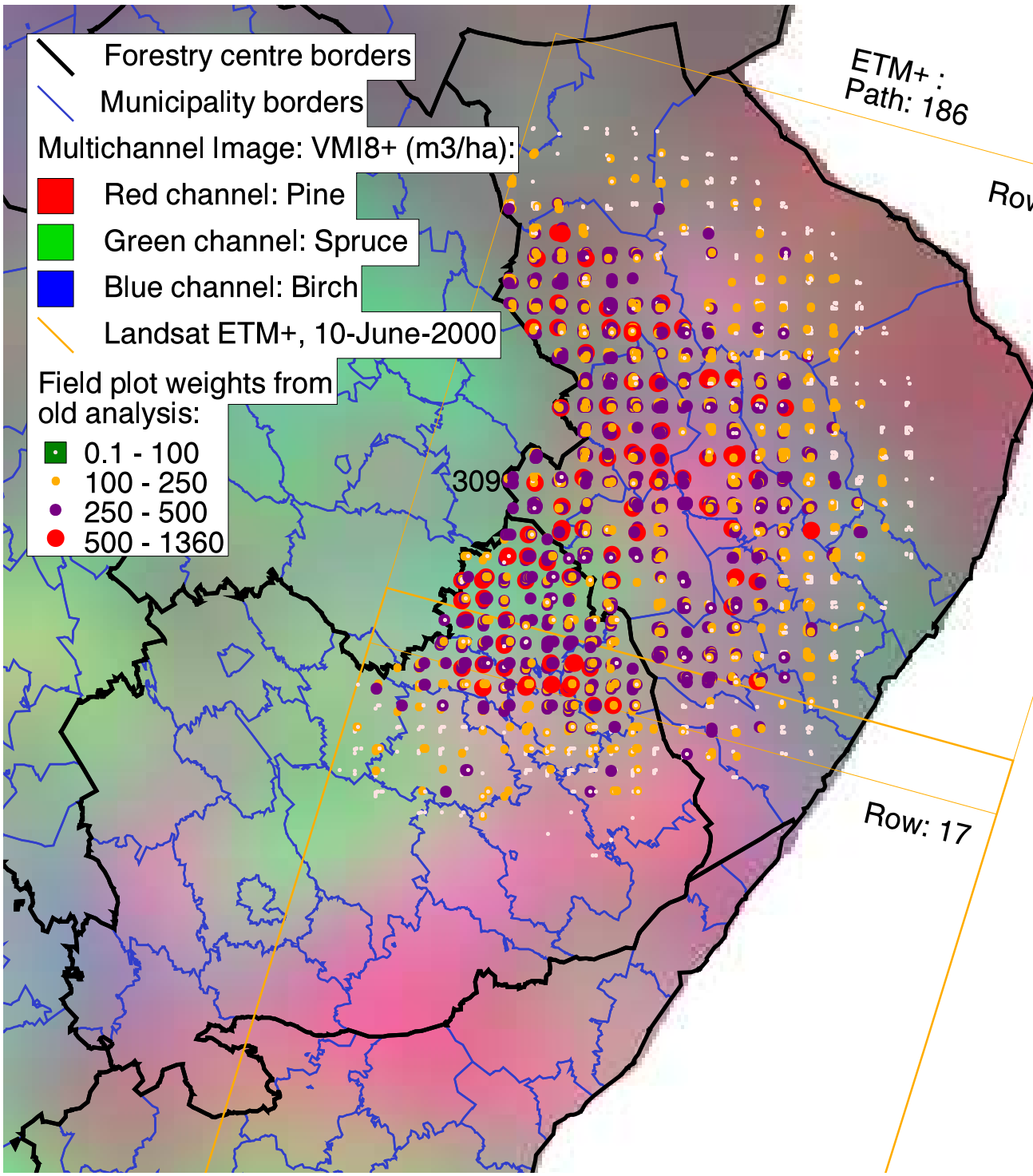
The fitness function to be minimised

$$f(\omega, \gamma, \hat{\delta}, \hat{\bar{e}}) = \sum_{i=1}^{n_e} \gamma_i \hat{\delta}_i(\omega) + \sum_{i=1}^{n_e} \gamma_{i+n_e} \hat{\bar{e}}_i(\omega)$$

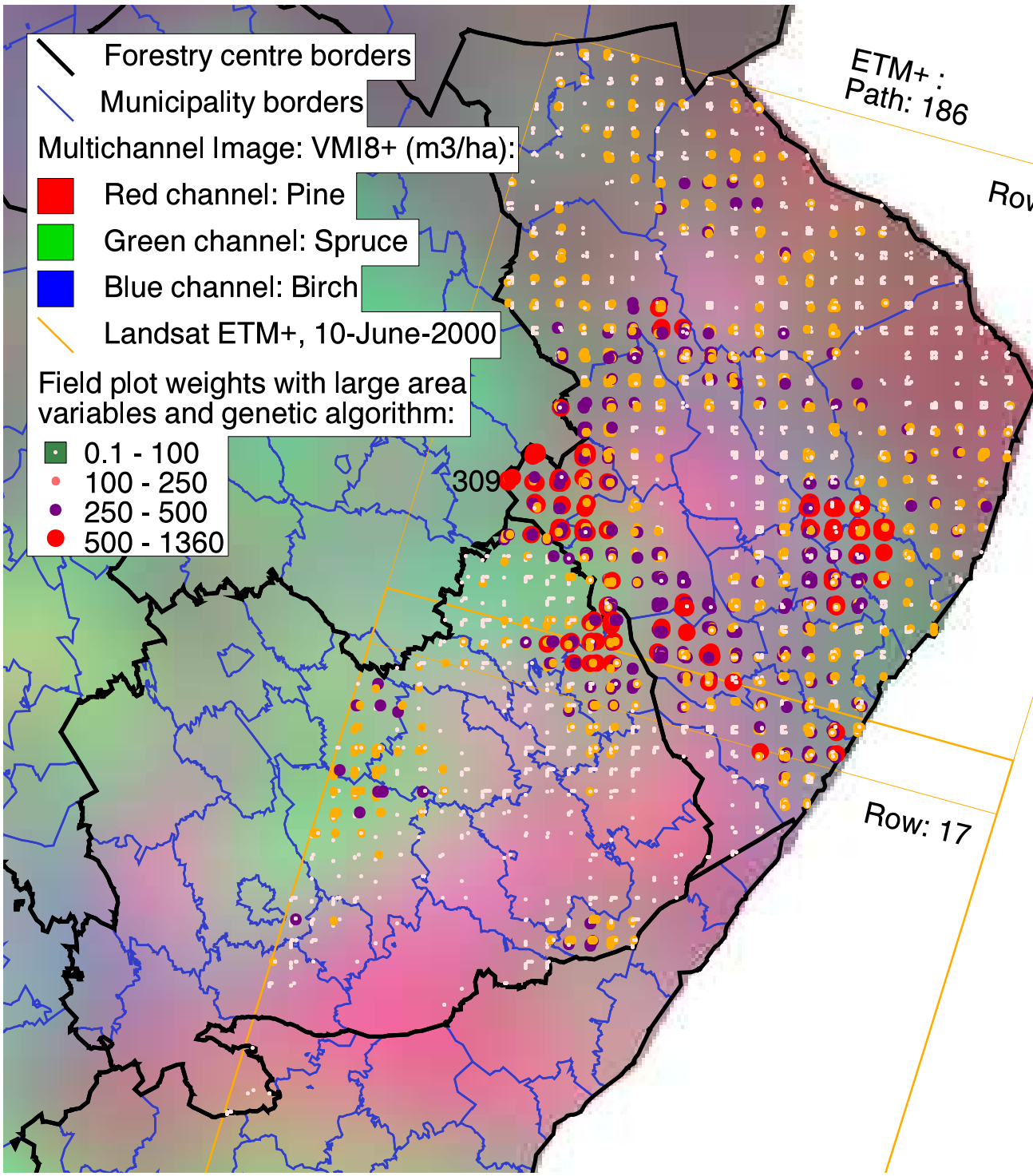
where

- $\hat{\delta}_j(\omega)$ is the standard error of the estimate of the forest variable i
- $\hat{\bar{e}}_j(\omega)$ is the bias of the estimate of the forest variable i
- n_e the number of the forest variables used in the algorithm
- γ_i fixed constants

Field plot weights for one municipality with the old method



Field plot weights for one municipality with the new method



Bias of the old and new k-nn estimates at pixel level (field plot level) using leave-one-out cross-validation and field data based volume estimates $\hat{V}_F$ in comparison, 1953 field plots.

	$\hat{V}_F$	Bias original	Bias new	Reduction
Volume	m^3/ha	m^3/ha	m^3/ha	%
Total	122.41	-2.871	-1.199	58.244
Pine	63.75	2.672	0.091	96.613
Spuce	38.88	-4.557	-0.567	87.565
Birch	15.90	-0.599	-0.481	19.611
O. br. l.	3.88	-0.448	-0.328	26.898

Difference between multi-source volume estimates $\hat{V}_M$ and field data based estimates $\hat{V}_F$ with old and new methods for municipality group b1, forest area 167 000 ha.

		old method	
Estimate of	$\hat{V}_M$ m^3/ha	$\hat{V}_M - \hat{V}_F$ m^3/ha	$(\hat{V}_M - \hat{V}_F)/\hat{\sigma}_F$
total volume	105.8	-6.9	-1.8
pine volume	54.7	-5.6	-1.3
spruce volume	29.7	0.4	0.1
birch volume	17.0	-0.8	-0.5
O. br. l.	4.4	-0.9	-0.9
weighted average of % bias		6.6	

		new method	
Estimate of	$\hat{V}_M$ m^3/ha	$\hat{V}_M - \hat{V}_F$ m^3/ha	$(\hat{V}_M - \hat{V}_F)/\hat{\sigma}_F$
total volume	111.5	-1.2	-0.3
pine volume	59.2	-1.1	-0.3
spruce volume	28.9	-0.4	-0.1
birch volume	17.6	-0.2	-0.1
O. br. l.	5.8	0.5	0.5
weighted average of % bias		2.0	

Difference between multi-source volume estimates $\hat{V}_M$ and field data based estimates $\hat{V}_F$ with old and new methods for municipality group IV, forest area 270 000 ha.

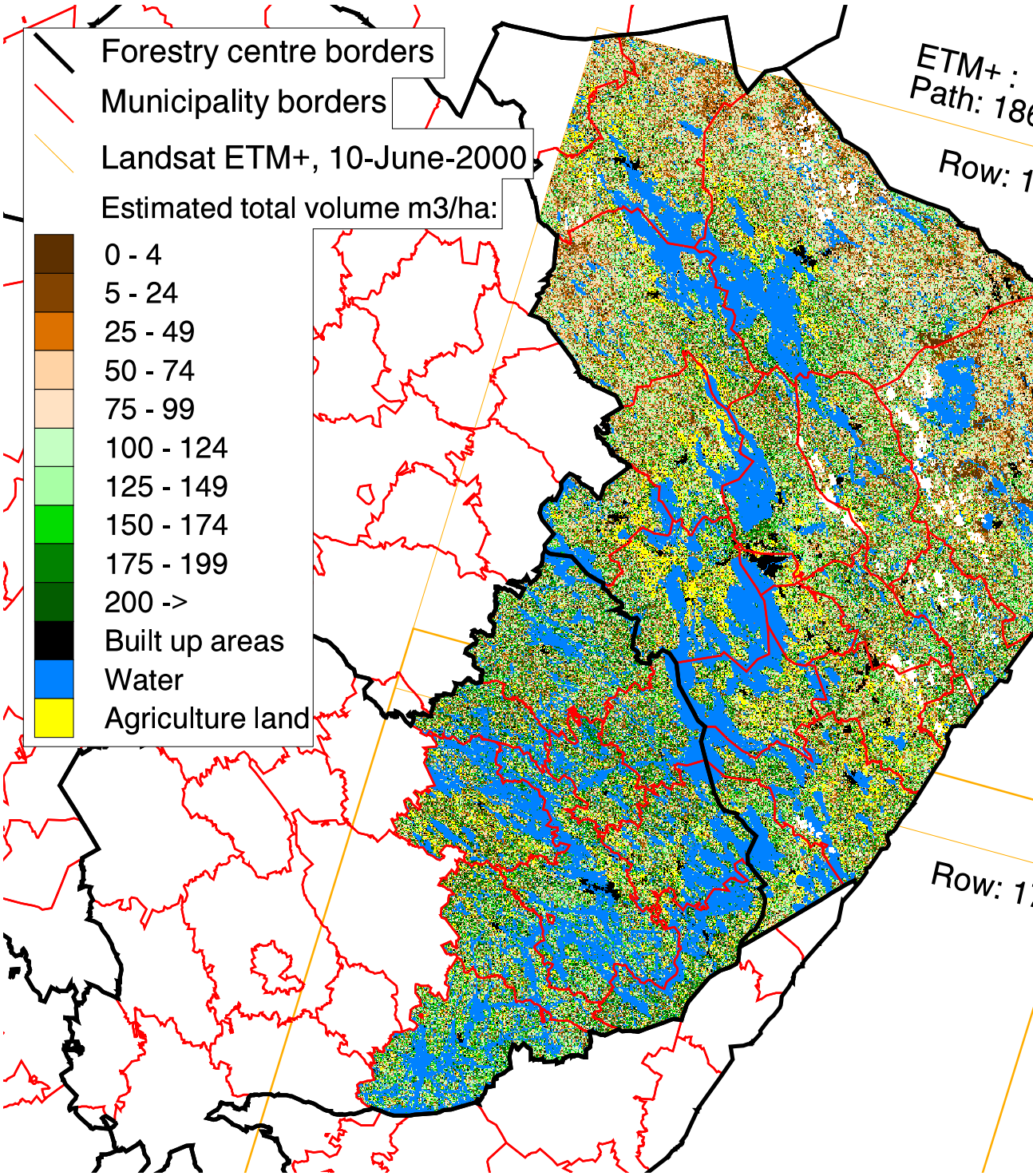
		old method	
Estimate of	$\hat{V}_M$	$\hat{V}_M - \hat{V}_F$	$(\hat{V}_M - \hat{V}_F)/\hat{\sigma}_F$
	m^3/ha	m^3/ha	
total volume	90.8	-8.8	-2.2
pine volume	48.4	3.4	1.4
spruce volume	25.3	-9.9	-3.4
birch volume	14.0	-1.7	-1.5
O. br. l.	3.1	-0.6	-1.0
weighted average of % bias		14.1	

		new method	
Estimate of	$\hat{V}_M$	$\hat{V}_M - \hat{V}_F$	$(\hat{V}_M - \hat{V}_F)/\hat{\sigma}_F$
	m^3/ha	m^3/ha	
total volume	90.9	-8.7	-2.2
pine volume	45.2	0.2	0.1
spruce volume	28.3	-6.9	-2.4
birch volume	14.3	-1.4	-1.3
O. br. l.	3.0	-0.7	-1.2
weighted average of % bias		8.4	

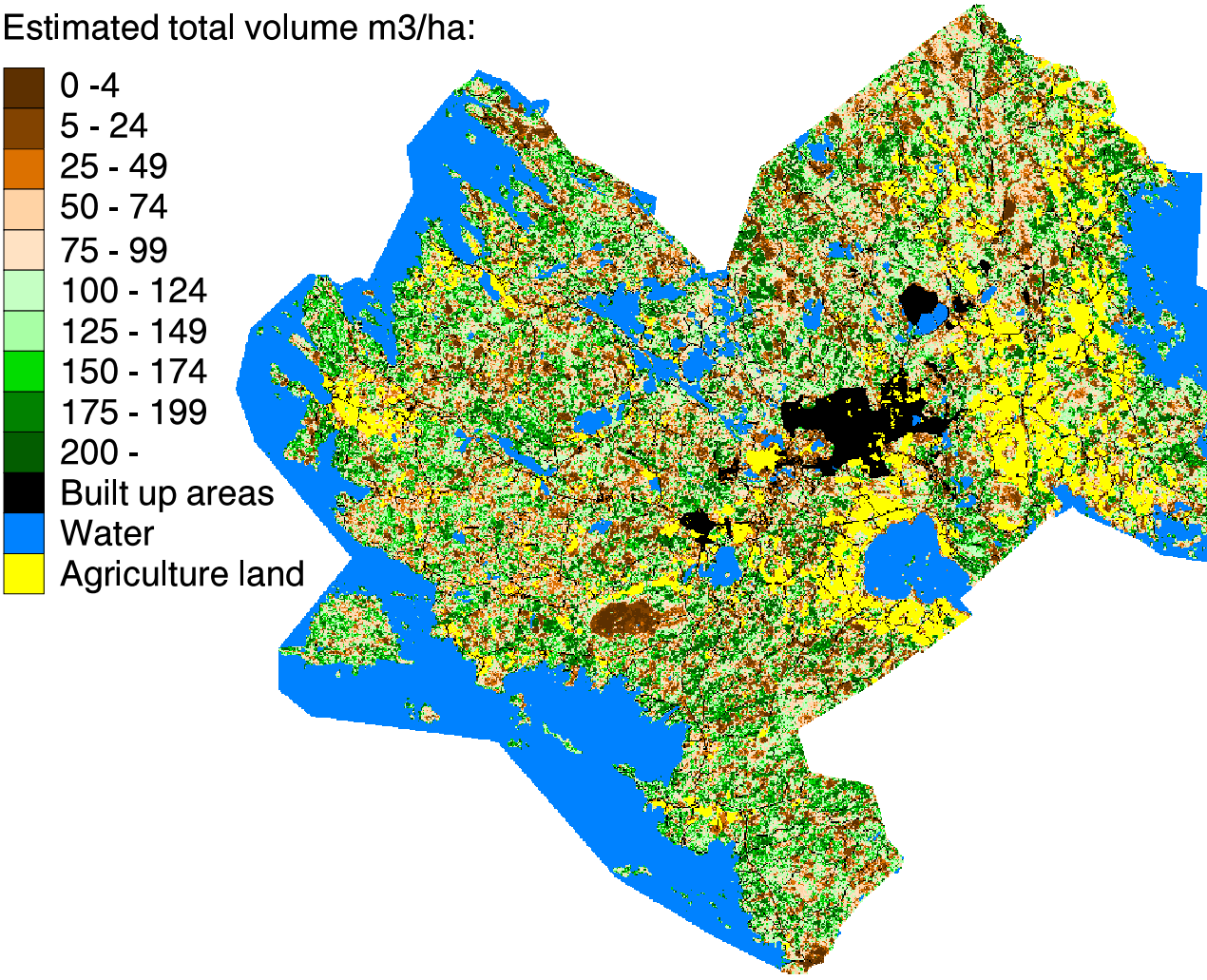
Difference between multi-source volume estimates $\hat{V}_M$ and field data based estimates $\hat{V}_F$ with old and new methods for North Karelia, forest area 1,49 mill. ha.

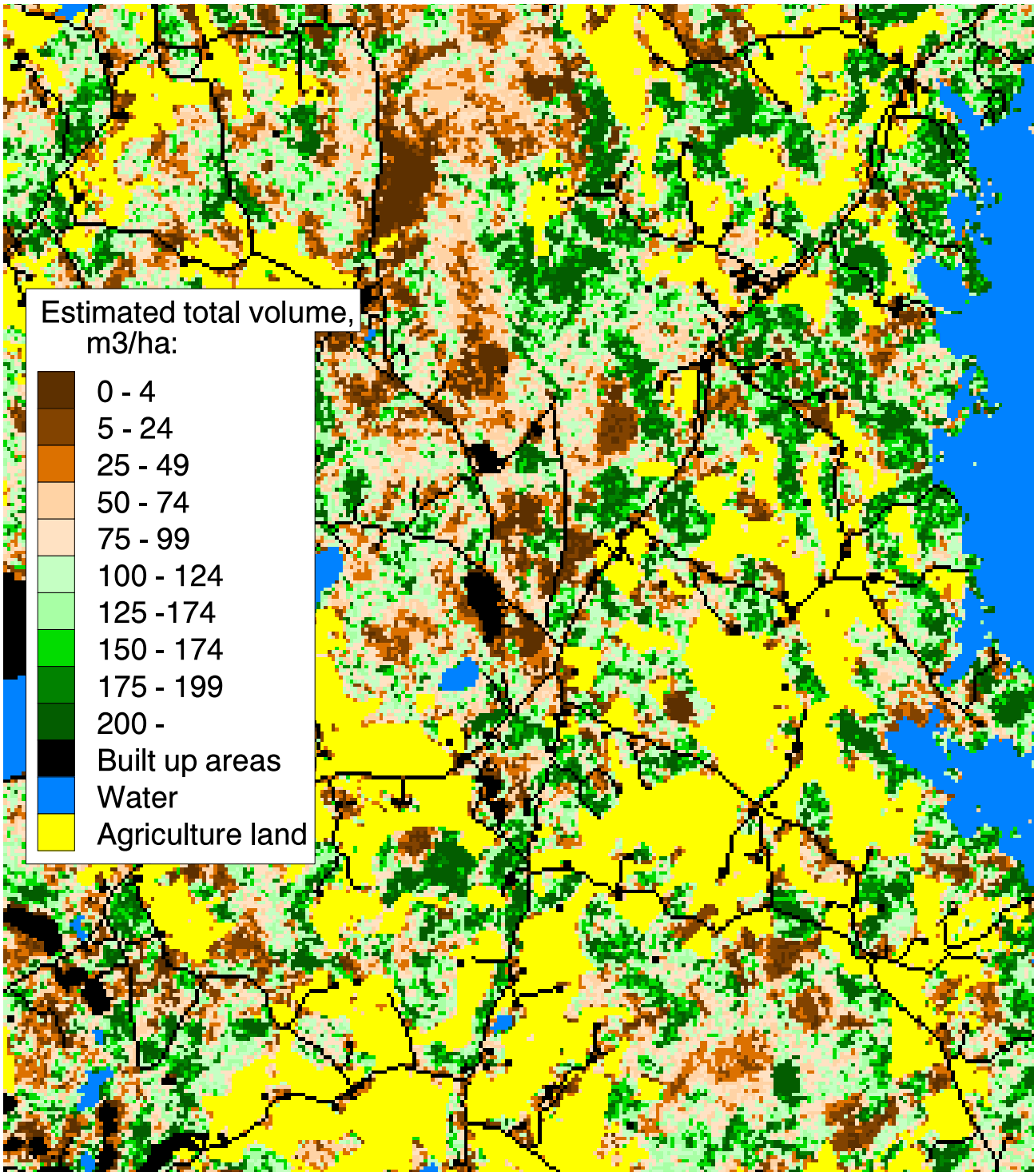
		old method	
Estimate of	$\hat{V}_M$ m^3/ha	$\hat{V}_M - \hat{V}_F$ m^3/ha	$(\hat{V}_M - \hat{V}_F)/\hat{\sigma}_F$
total volume	101.6	-1.3	-0.9
pine volume	54.1	1.6	1.3
spruce volume	29.3	-1.2	-1.1
birch volume	14.9	-0.9	-1.8
O. br. l.	3.3	-0.7	-2.3
weighted average of % bias		4.2	

		new method	
Estimate of	$\hat{V}_M$ m^3/ha	$\hat{V}_M - \hat{V}_F$ m^3/ha	$(\hat{V}_M - \hat{V}_F)/\hat{\sigma}_F$
total volume	101.4	-1.5	-1.0
pine volume	52.6	0.1	0.1
spruce volume	30.4	-0.1	-0.1
birch volume	14.8	-1.0	-2.0
O. br. l.	3.5	-0.5	-1.7
weighted average of % bias		1.6	


Forestry centre borders
Municipality borders
Landsat ETM+, 10-June-2000
Estimated total volume m3/ha:
0 - 4
5 - 24
25 - 49
50 - 74
75 - 99
100 - 124
125 - 149
150 - 174
175 - 199
200 ->
Built up areas
Water
Agriculture land
ETM+ :
Path: 18
Row: 1
Row: 1

Outokumpu municipality




Estimated total volume, m3/ha:
0 - 4
5 - 24
25 - 49
50 - 74
75 - 99
100 - 124
125 -174
150 - 174
175 - 199
200 -
Built up areas
Water
Agriculture land

— Properties of k-nn estimates —

- COST-EFFICIENCY: much more detailed information with very low additional costs compared to FIELD MEASUREMENT based method
- SYNTHETIC ESTIMATOR: information outside the area is utilised
- NON-PARAMETRIC: no need for implicit modelling of the dependence between ground and image data
- PRACTICAL: suitable for
  - all inventory parameters can be estimated for each computation unit
  - thematic maps and small area statistics
- STATISTICALLY ORIENTED: preserves natural dependence structure between forest variables
- VERSATILE: same method can be applied
  - to very different environments
  - to different types of remote sensing material

— Properties of k-nn estimates, continuation —

- Pixel level RMSE of the estimates by cross validation
- $\hat{v} - v$ depends on $v \Rightarrow \hat{v} - v$ spatially correlated
- Spatial dependencies in the image itself make the error structure complex
- Several error sources
  - inaccuracy in field plot locations
  - a field plot does not represent a pixel
  - the spectral variation does not explain the variation of the field data vector
  - between image variation in image quality, etc.
- Cross validation may underestimate the actual pixel level error, if the variation in the field plot data does not cover all of the actual variation in the forest area in question
- Developing an operationally usable statistical error assessment technique is a highly challenging task and a fully satisfactory solution is yet to be found

— Conclusions, ancillary information and genetic algorithm —

$\Longrightarrow$ large scale forest variable estimates as additional variables reduce biases

$\Longrightarrow$ the variable weight selection for image variables and large scale forest variable estimates

- difficult optimisation problem, the fitness function is highly nonlinear and non-continuous with respect to weights

$\Longrightarrow$ fitness function unexpectedly flat, the introduction of upper bounds for the weight restricted the space under search (no big thing as for the final solution)

$\Longrightarrow$ numerous versions of the model were dealt with, incredible amount of CPU time was consumed - with metaheuristics normally good solutions are achieved with the cost of great deal of fine-tuning

$\Longrightarrow$ the biases could be reduced and the accuracy of the estimates improved

- both new 'search space' and weights affect

$\Longrightarrow$ in an operative use

$\Longrightarrow$ details will be given in a forthcoming paper

— Area and volume estimates from field data - continuation —

Instead of global variation, $(z_i - \bar{z})^2$, Matérn considers the local variation.
Define z_i = $x_i - m_n y_i$ for a group g of four clusters

$$i1 \quad i2$$

$$i3 \quad i4$$

The local variation of the group g is described by a quadratic form

$T_g = \frac{1}{4}(z_{i1} - z_{i2} - z_{i3} + z_{i4})^2$.

and error by

$s = \frac{\sqrt{\sum_g T_g}}{\sum_i^n y_i}$, where g goes over the cluster groups.

Matérn defines the asymptotic error variance per sample point as

$$\sigma^2(m_n) = \lim_{n\to\infty} nE\{[\overline{z}_n - z(A_n)]^2\},$$

$A_n = (na)^2$, a being a fixed number, can be specified for example as a square of area.

1. Initialisation Generate the initial population with npop elements of random weight vectors.

2. Tournament selection Choose among the current population of weight vectors a random pair of vectors the fitness values of which are compared. Roughly speaking - the better is chosen to be part of the population continuing to crossover. Repeat until the population consists of npop vectors.

3. Crossover With two vectors a and b (parent vectors) carry out uniform crossover, i.e. with probability cross take the ith element from a and with probability (1-cross) from b. Thus two new vectors (children) emerge. Pick the best vector in fitness among children and parents as member of the next generation.

4. Mutation In each vector each element is mutated with probability muta. Two kinds of mutations can occur: radical (element is deducted from 1) or non-radical (element change 10 per cent). If the mutated vector's fitness is less than the original vector's fitness, it replaces the original one with probability. If less than ngen generations have been dealt with, go to 2, otherwise stop.

— The simplified algorithm, **to be checked and completed** —

1. Generate a set of random weights
   $V \subset W$, $\mathbf{v}^i \in R^n$, $(\mathbf{v}^i, i = 1, ..., npop)$, where
   W is the feasible set of the weight vectors ω, satisfying $0 \leq \omega_k \leq uppe_k$, $\|\omega\| = 1$.
2. Tournnament
   repeat $i = 1, ...nppop$.
   For randomly picked up $\mathbf{v}^a$ and $\mathbf{v}^b$:
   if $f(\mathbf{v}^a) > f(\mathbf{v}^b)$, then $\omega^i \leftarrow \mathbf{v}^a$ with P = $tp1$
   $\omega^i \leftarrow \mathbf{v}^b$.
   If the former happens, then with P = $tp2$, $\omega^{i+1} \leftarrow \mathbf{v}^b$, otherwise $\omega^{i+1} \leftarrow \mathbf{v}^b$.
3. Crossover
   repeat for pairs ω^i and ω^{i+1} $i = 1, ...npop$ an uniform crossover with P = $erpr$.
4. Mutation
   repeat $i = 1, ...nppop$ and $k = 1, ..., n$: mutate element ω^i_k with P = $muta$ and and radical mutate with P = $radi$.
5. Repeat 2 - 4 $ngen$ times (= number of generation).

— The parameter selection, **to be completed** —

Through trial and error using

- cross-validation
- to some extent estimates for groups of municipalities and their standard errors computed from field data

The parameters
$muta$ = 0.05
$radi$ = 0.35
$erpr$ = 0.75.

performed well.

Difference between multi-source volume estimates $\hat{M}$ and field data based estimates $\hat{F}$ with old and new methods for municipality group 1, forest area 167 000 ha.

		old method	
	m^3/ha	m^3/ha	
Estimate of	$\hat{M}$	$\hat{M}-\hat{F}$	$(\hat{M}-\hat{F})/\hat{\sigma}_F$
total volume	105.8	-6.9	-1.8
pine volume	54.7	-5.6	-1.3
spruce volume	29.7	0.4	0.1
birch volume	17.0	-0.8	-0.5
o. br. l.	4.4	-0.9	-0.9
weighted average of % bias		6.6	

		new method	
	m^3/ha	m^3/ha	
Estimate of	$\hat{M}$	$\hat{M}-\hat{F}$	$(\hat{M}-\hat{F})/\hat{\sigma}_F$
total volume	114.9	2.2	0.6
pine volume	62.3	2.0	0.5
spruce volume	28.9	-0.4	-0.1
birch volume	18.4	0.6	0.4
o. br. l.	5.3	0.0	0.0
weighted average of % bias		2.7	

Difference between multi-source volume estimates $\hat{M}$ and field data based estimates $\hat{F}$ with old and new methods for municipality group 4, forest area 270 000 ha.

		old method	
	m^3/ha	m^3/ha	
Estimate of	$\hat{M}$	$\hat{M}-\hat{F}$	$(\hat{M}-\hat{F})/\hat{\sigma}_F$
total volume	90.8	-8.8	-2.2
pine volume	48.4	3.4	1.4
spruce volume	25.3	-9.9	-3.4
birch volume	14.0	-1.7	-1.5
o. br. l.	3.1	-0.6	-1.0
weighted average of % bias		14.1	

		new method	
	m^3/ha	m^3/ha	
Estimate of	$\hat{M}$	$\hat{M}-\hat{F}$	$(\hat{M}-\hat{F})/\hat{\sigma}_F$
total volume	95.6	-4.0	-1.0
pine volume	46.5	1.5	0.6
spruce volume	31.1	-4.1	-1.4
birch volume	14.2	-1.5	-1.4
o. br. l.	3.7	0.0	0.0
weighted average of % bias		6.8	