Utilizing prior information in environmental inventory design - experiences from forest inventories

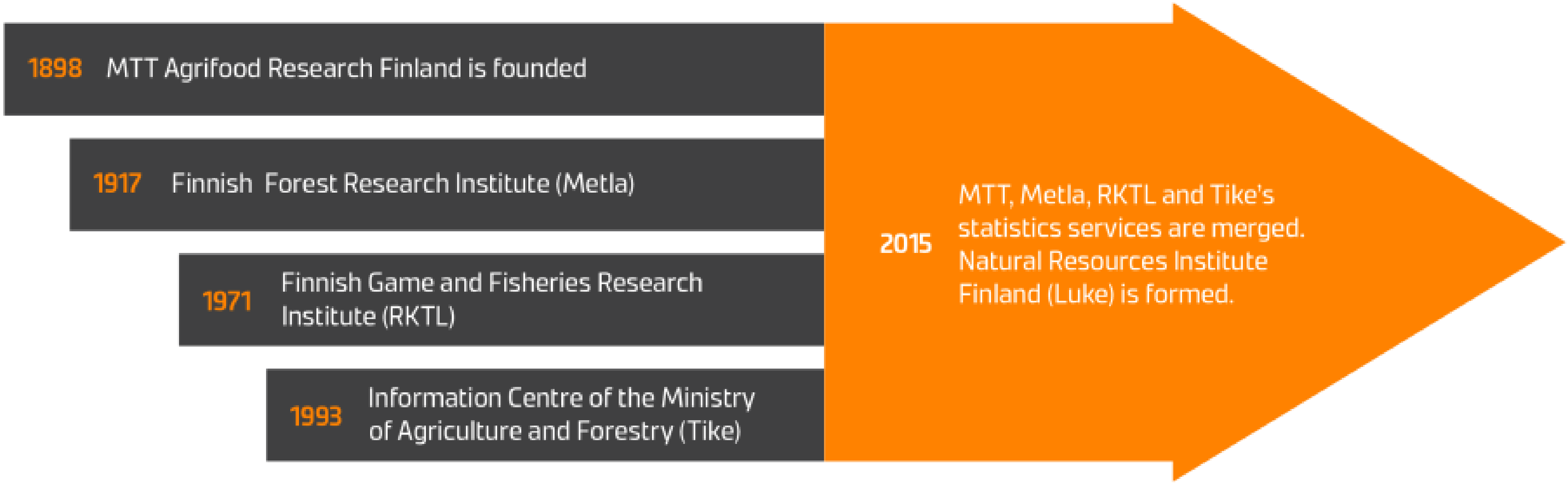
Juha Heikkinen, Pekka Hyvönen, and Helena M. Henttonen

XXVIIIth International Biometric Conference.

Victoria, Canada, July 10, 2016

NATURAL RESOURCES INSTITUTE FINLAND

Natural Resources Institute Finland (Luke)



Directorate 23
Scientists 669
Other experts 314
Research support personnel 431

December 2015

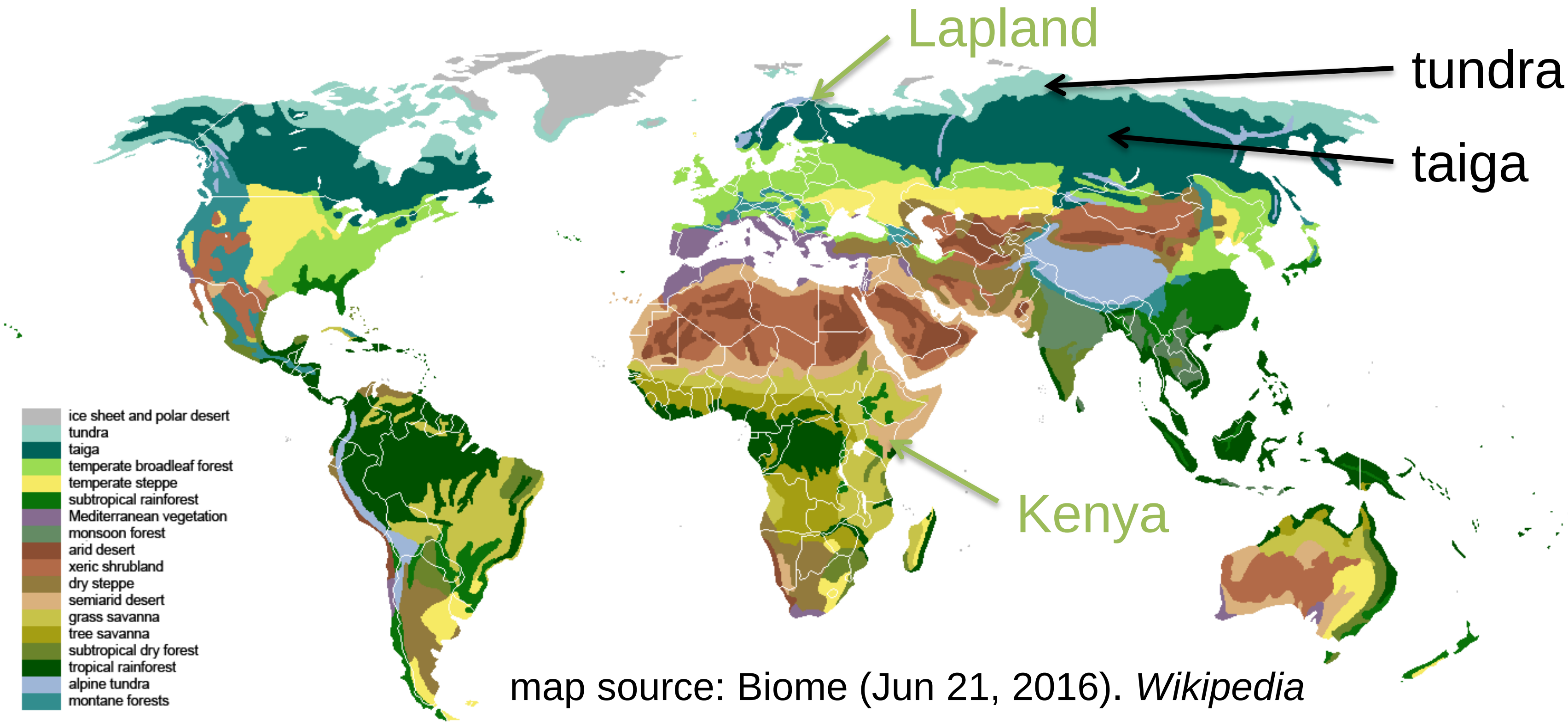
Among Luke's activities

- National Forest Inventory (NFI) of Finland
- Land use, land-use change and forestry (LULUCF) sector of national greenhouse gas inventory
- International consultation in inventory design and related capacity building
  - Tanzania (Tomppo et al. *Can. J. For. Res.* 2014)
  - Vietnam
  - Nepal
  - Cambodia
  - Kenya

Study areas

Nakuru, Kenya: highly fragmented forests, 6% of area

Lapland, Finland: borderline between taiga (Boreal forest) and tundra



map source: Biome (Jun 21, 2016). *Wikipedia*

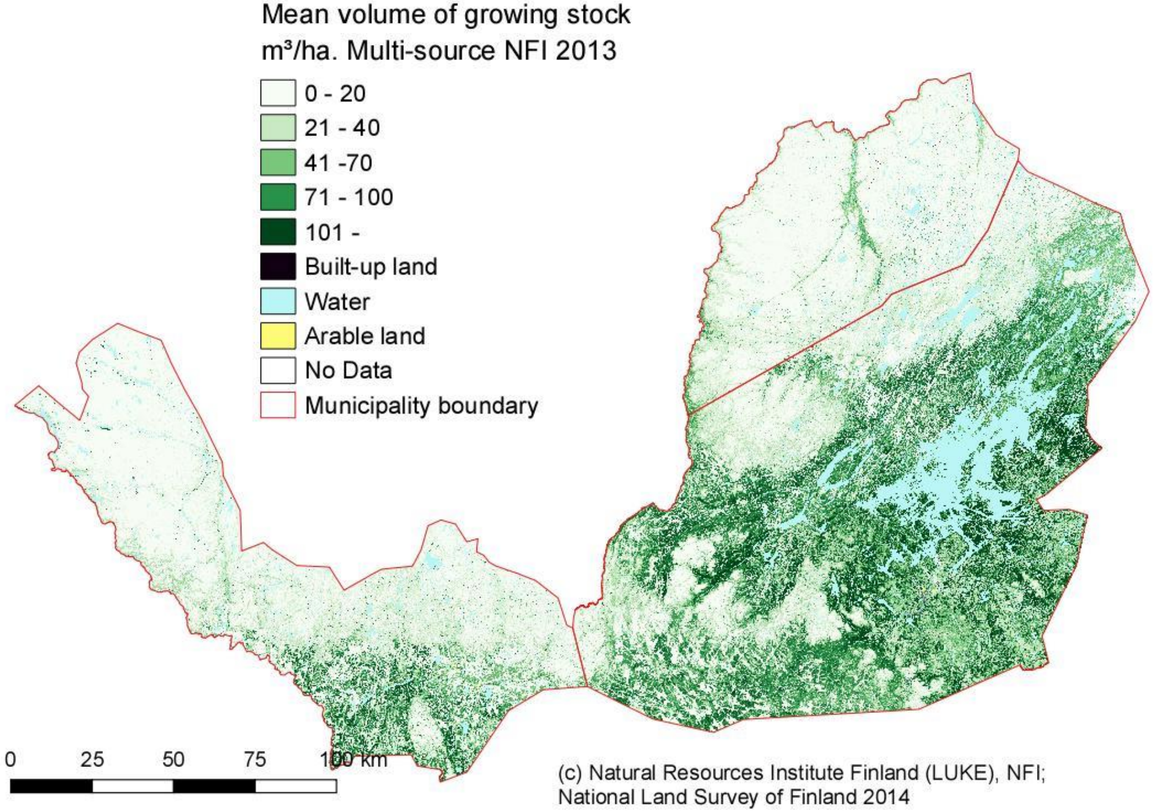
Luke NATURAL RESOURCES INSTITUTE FINLAND

Inventories of natural resources and the environment

- General aim: estimate total or mean of resource over (often administrative) region of intererst.
- Example: mean tree biomass by species in forests of Kenya.
- Unbiased, precise, and timely estimation: substantial amount of field measurements.
- Often fieldwork can be wisely targeted by using prior information.
- Example: thematic maps from similar or related earlier inventories.

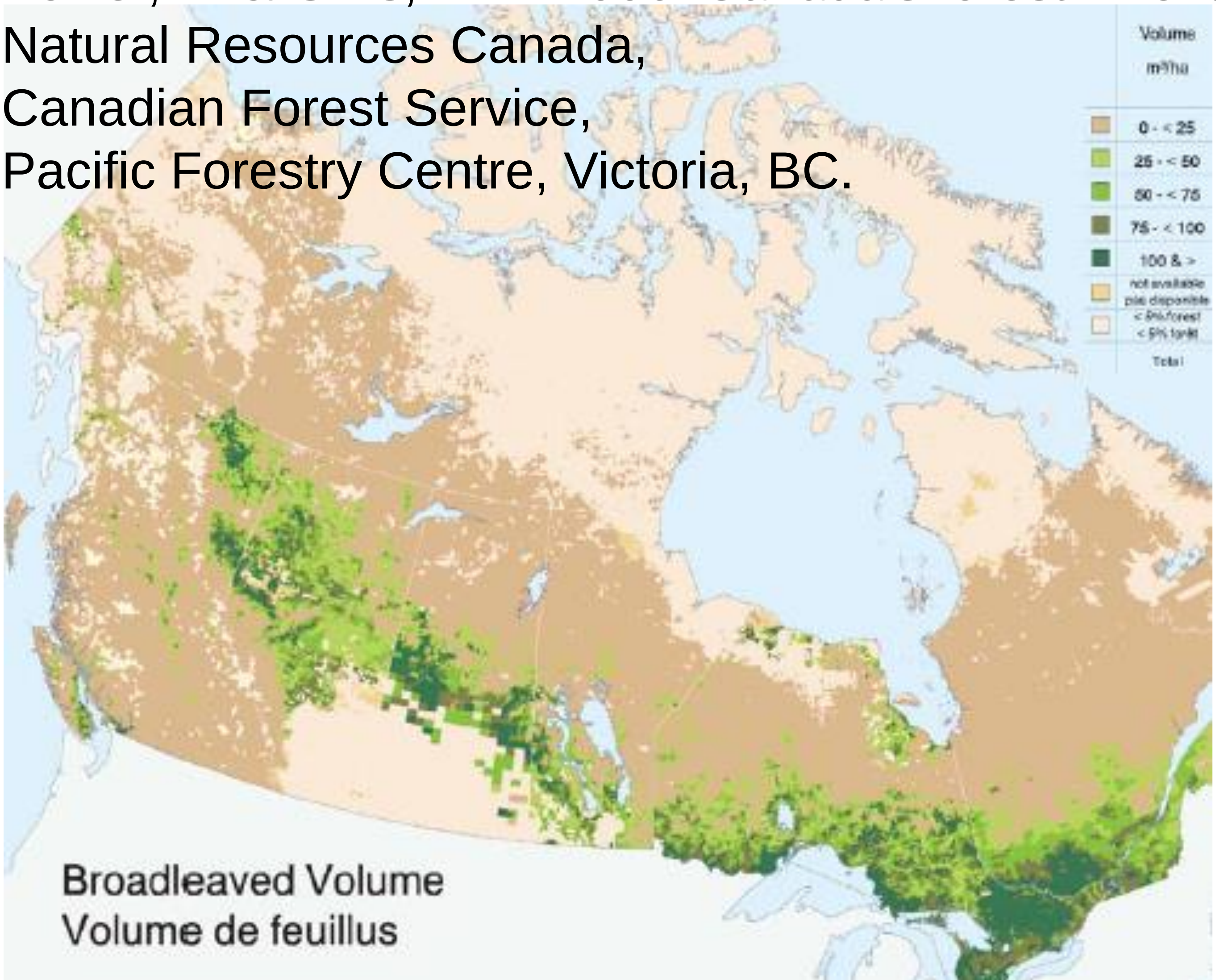
Example of prior information

Tree stem volume in Northernmost Lapland (for methods, see Tomppo et al. *Multi-Source National Forest Inventory*. Springer 2008).



Another example of prior information

Power, K. & Gillis, M.D. 2006. *Canada's forest inventory 2001*. Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Victoria, BC.



Luke NATURAL RESOURCES INSTITUTE FINLAND

Question

When estimating
$$\bar{Y} = \frac{1}{|A|}\int_A y(s)ds$$

where $A \subset \mathrm{R}^2$ region and *y* response surface of interest, e.g.

- $y(s) = \mathbf{1}_F(s) \Rightarrow \bar{Y}|A|$ area of *F*
- $y(s)$ mean biomass in small plot around $s \Rightarrow \bar{Y}$ mean biomass over *A*

by
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

where $y_i = y(s_i)$ field observations of *y* at *n* samle points s_i,

how can we use prior information related to variability of *y* available at all $s \in A$, when choosing the sampling locations s_i.

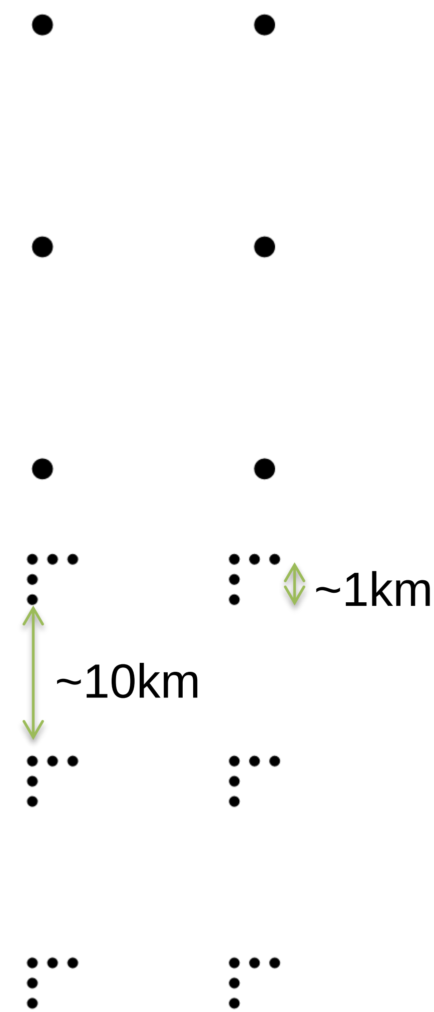
Common designs without prior information

Systematic plot sampling: e.g., square grid of sample plot centers for field measurements

- “regularly spaced design points are optimal for a variety of reasonable spatial correlation functions” (Stevens and Olsen, *J. Amer. Statist. Assoc.* 2004).

Systematic cluster sampling (sys1): e.g., square grid of sample plot clusters

- Large-scale inventory: one grid location/day.
- Distribute one day’s work between plots in different neighbouring forest stands rather than measure a large number of similar trees from one stand.



Stratified sampling and Neyman allocation

Example (unrealistically simplified setting for illustration):
Use prior information to

- divide inventory region A into two **strata** A_h so that variation of y high in A_1 and low in A_2, (e.g. map of forest types) and to
- estimate variances S_h^2 of y within strata $h = 1, 2$ (e.g. biomass map from earlier inventory).

Neyman allocation: sampling densities $p_h = n_h \,/\, |A_h|$ determined by $p_1 / p_2 = S_1 / S_2$, where

- n_h is the sample size and
- $|A_h|$ the area of stratum h.

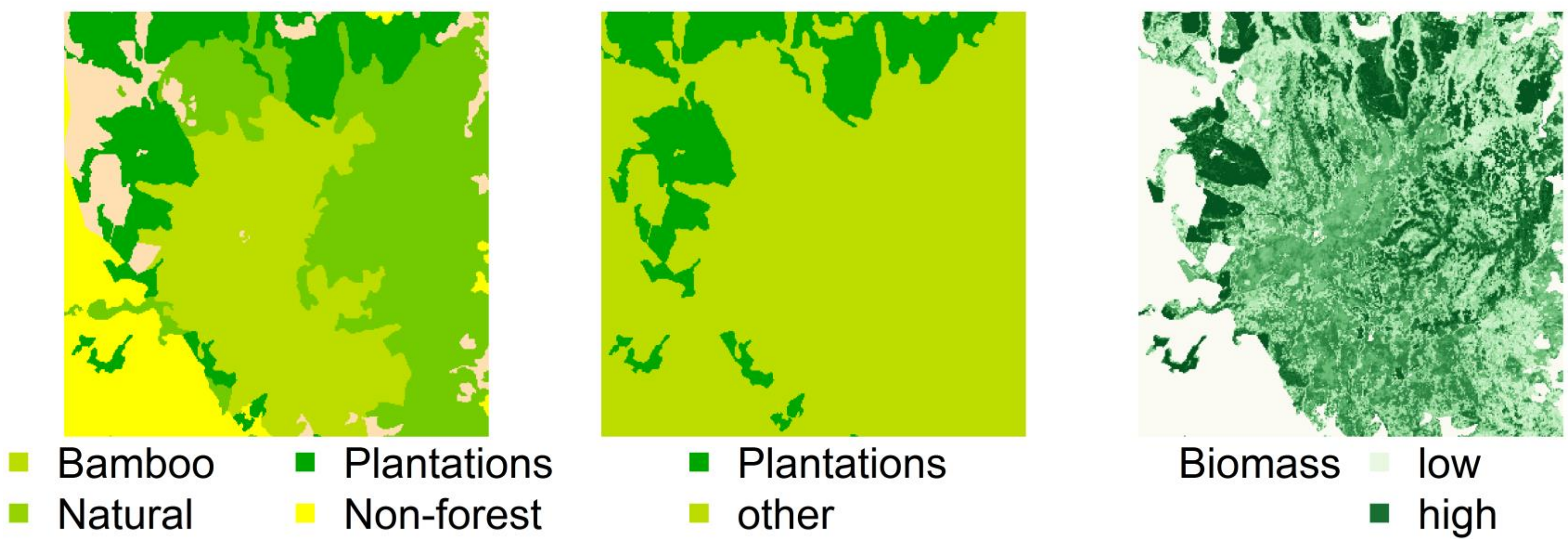
Example: mean tree biomass in Kenya

Vegetation classes (left) ⇒ stratification (middle):

- **Stratum 1: plantations** (high mean biomass -> high variability), **2: other forests and non-forest**

Biomass in forest (right) ⇒ Stratum variances S_h^2

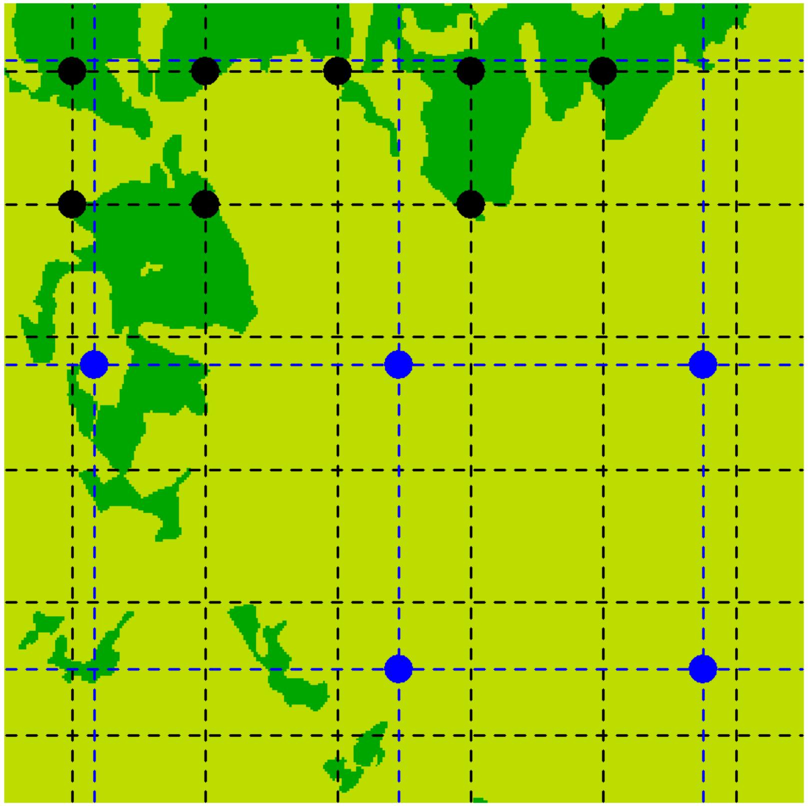
Note: if stratum h non-forests, then $S_h = 0$



Stratified systematic plot sampling

Two grids with densities d_h, such that $d_1/d_2 = S_1/S_2$.

Sample from stratum h: $d_h \cap A_h \Rightarrow p_1/p_2 \approx S_1/S_2$



Stratum 1: plantation forests

(high mean biomass, high variability)

Stratum 2: other forests and non-forest (low mean biomass, low variability)

Luke NATURAL RESOURCES INSTITUTE FINLAND

Stratified systematic cluster sampling

Variable number of plots / cluster

$\Rightarrow$ idea of one day per cluster is lost.

Example: one cluster from **stratum 1 grid**; only 2 plots within **stratum 1** included in the sample.

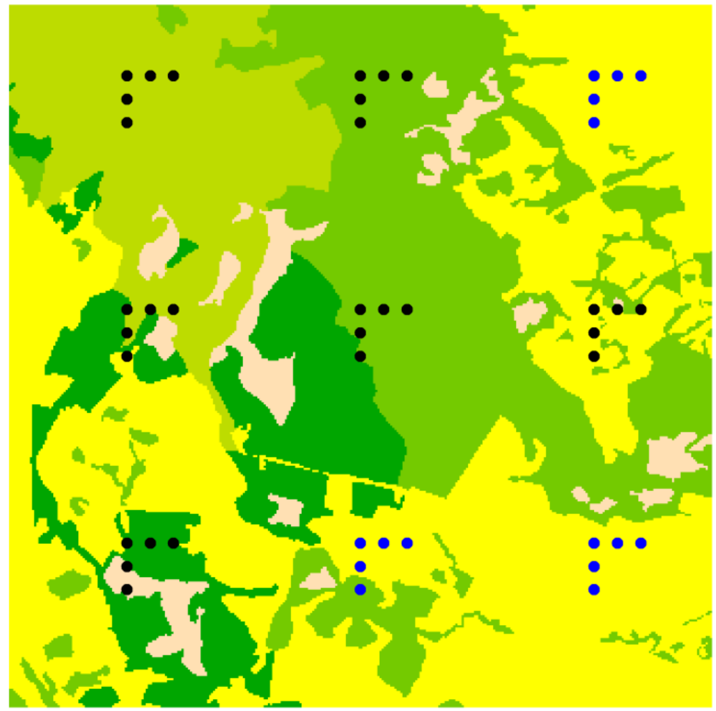
Two-phase (double) sampling

Instead of single points, stratify whole clusters.

Stratum 1: clusters with 2 or more forest plots

Stratum 2: clusters with 0 or 1 forest plots

Note: No need to specify high/low-biomass forests



- forest
- forest
- forest
- non-forest
- non-forest

First phase sample must be dense

In simple (one-phase) stratified sampling

- exact stratum **weights** for stratified estimation from $|A_h|$
- and **variances** S_h between clusters for Neyman allocation from auxiliary data at population level.

In double sampling (of clusters) weights and variances are estimated from 1st-phase sample, e.g., dense systematic grid.

2nd-phase sample, measured in field. Often, but not necessarily, subsample of 1st-phase sample.

Three strategies for 2nd-phase sampling

ran2: simple random subsamples of n_h clusters from 1st-phase sample

- dense 1st phase for good weights & to enable uneven allocation $\Rightarrow$ low sampling fraction, esp. in stratum 2 $\Rightarrow$ spatial balance lost
- fixed sample size possible $\Rightarrow$ exact Neyman allocation

sys2: as earlier for stratified systematic plot sampling

- new grids with densities according to Neyman allocation
- optimal spatial balance within strata
- random n_h: only approximate Neyman allocation

bal2: Grafström & Tillé (*Environmetrics* 2012)

- stratified spatially balanced subsampling from 1st-phase sample
- fixed sample size possible for each $h \Rightarrow$ exact Neyman allocation
- spatial balance also across strata, but sub-optimal (within strata) in comparison to systematic sampling

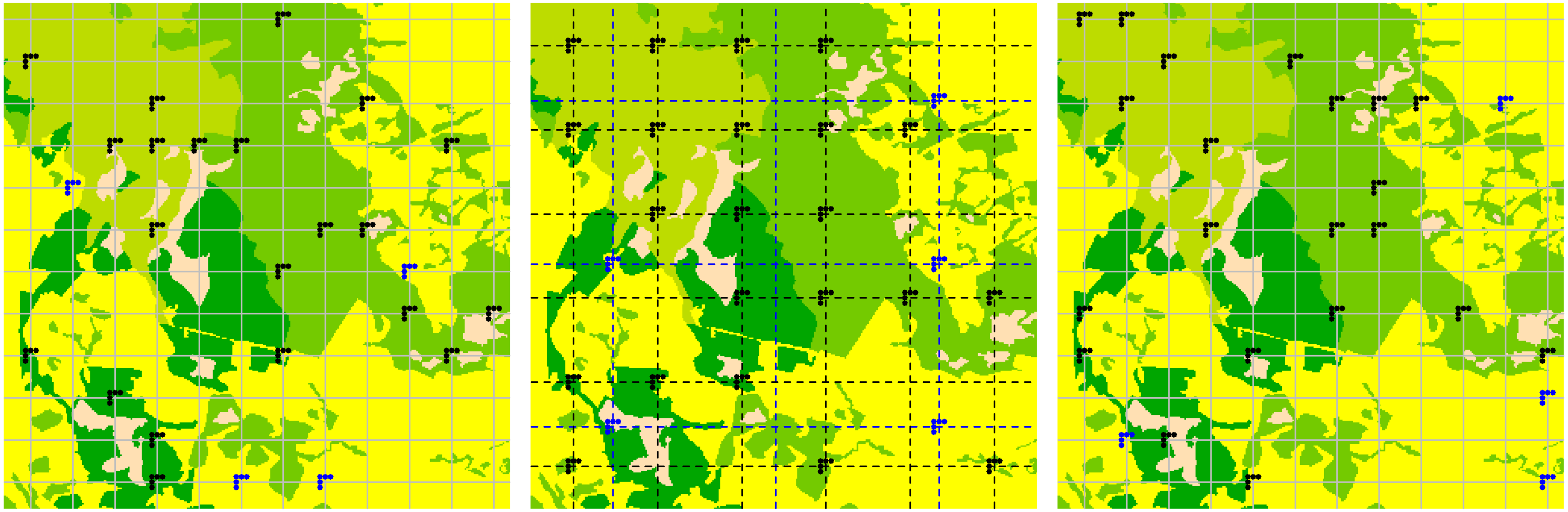
Stratified two-phase cluster samples

Target allocation

Stratum 1: 21 clusters

Stratum 2: 4 clusters

gray grid: 1st-phase sample



ran2 sys2 bal2

Luke NATURAL RESOURCES INSTITUTE FINLAND

Comparison of designs: sampling simulation

Anticipated variance of inventory estimator by

- simulating replications of each design
- comparing sample means to known population mean of biomass map
- estimating variance by MSE over replications

This will typically underestimate the variance of the actual inventory estimator, because some of the natural variation is smoothed out in multi-source maps.

However, if spatial structure of biomass map reflects true spatial structure in scales that determine cluster-to-cluster variation, then such sampling simulation should be useful in comparison of different sampling designs.

Sampling simulator developed at Metla/Luke

Also allows comparison of different cluster forms and can take into account transfer time between sample plots

- digital terrain model, vegetation type etc.

and measurement time

- biomass or vegetation type

Has been applied in

- Vietnam 2012-2013
- Cambodia 2014
- Kenya 2015

More information:

pekka.hyvonen@luke.fi

kari.t.korhonen@luke.fi

Case studies on three test areas

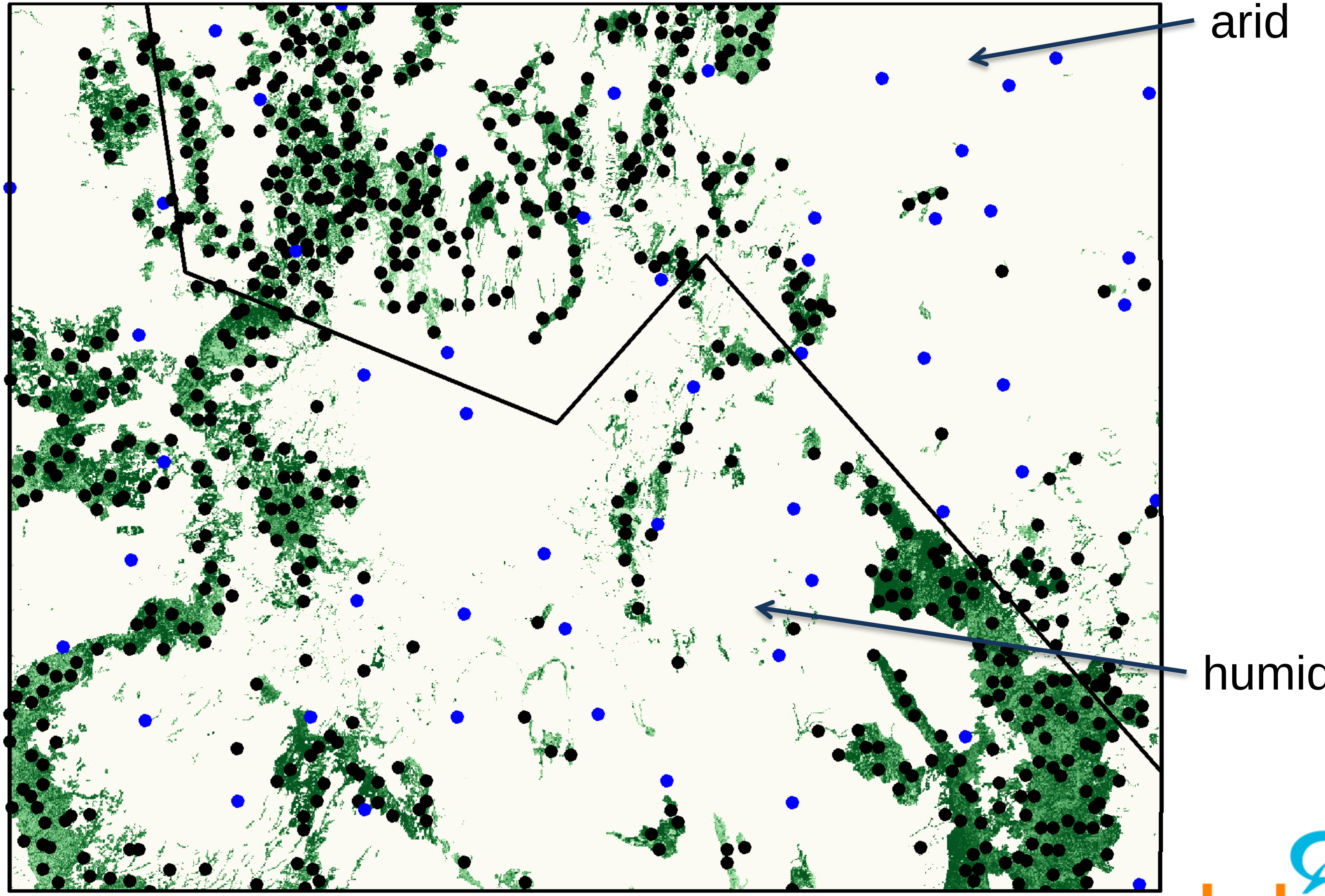
	$\lvert A \rvert$ km^2	% forest	$S/\bar{y}$	n clusters	plots/ cluster
Lapland	28,140	0.99	1.0	42	9
Nakuru, arid	9,270	0.16	3.2	376	5
Nakuru, humid	13,590	0.23	2.6	374	5

Qualitative differences

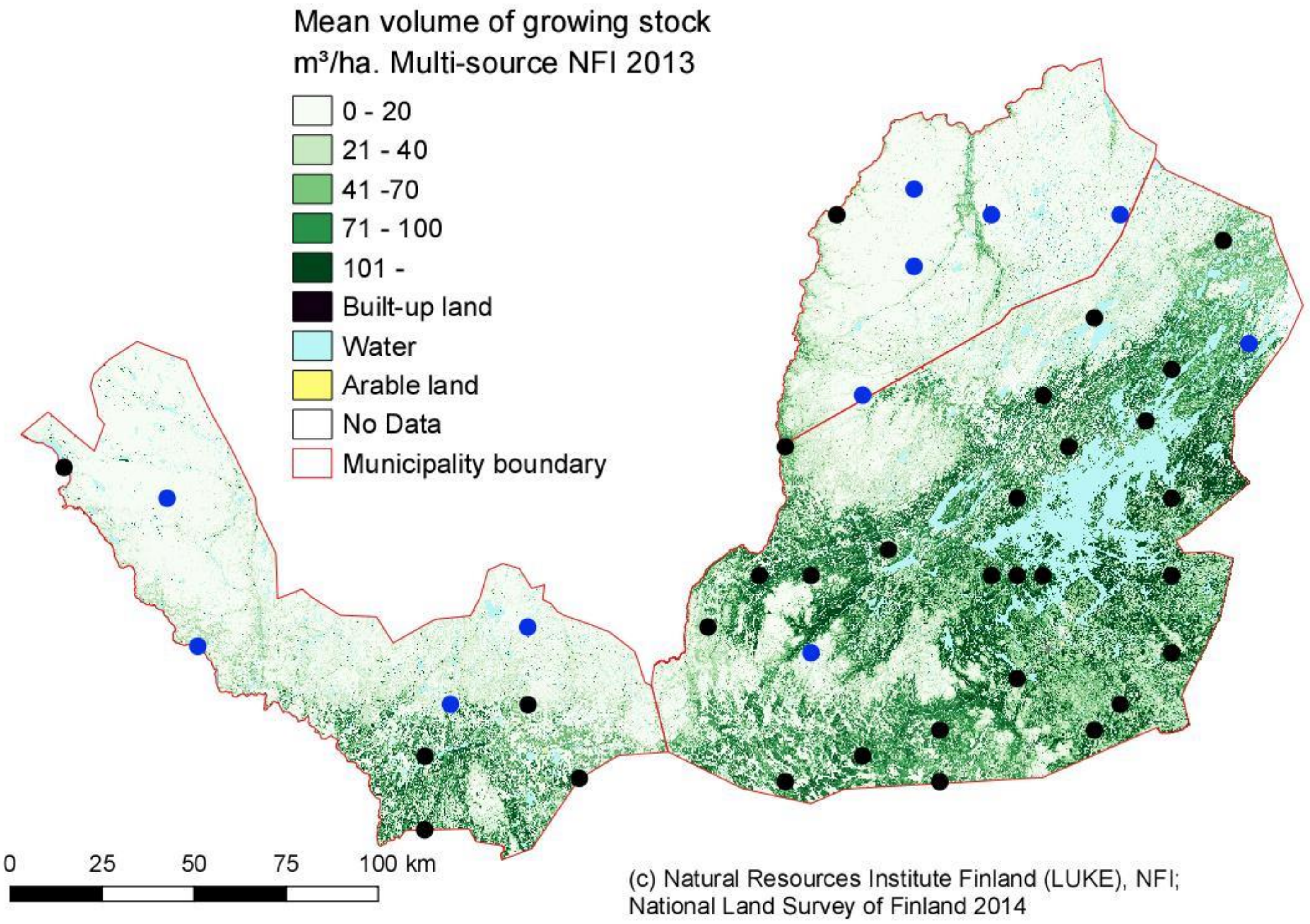
- Forests more concentrated in Lapland (SE), scattered in Nakuru
- Less plots/cluster in Nakuru due to more measurements/plot, more difficult movement, shorter day
- Variability of *y* different in all three areas

Target: 10% relative sampling error by **sys1**

Stratified two-phase bal2-design: Nakuru



Stratified two-phase bal2-design: Lapland

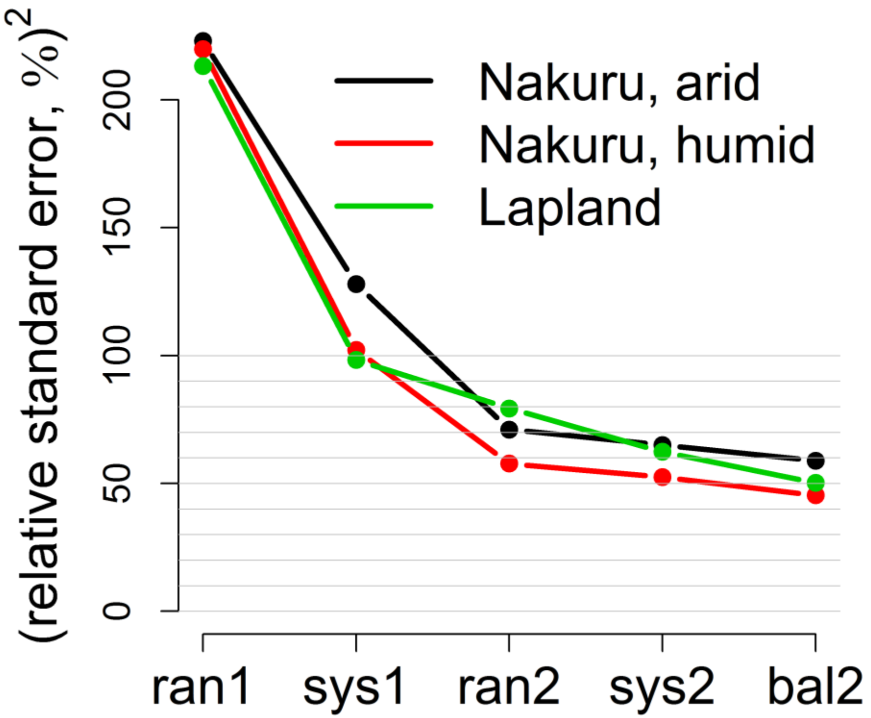


Preliminary results

ran1: 1-phase simple random sampling

Why RSE²? Ratio between two designs gives the ratio between *n* yielding same precision.

Below: Equivalent sample sizes within each row.



	ran1	sys1	ran2	sys2	bal2
Lapland	424	196	158	124	100
Nakuru, arid	379	217	121	110	100
Nakuru, humid	485	225	127	116	100

Conclusions

- Clustered designs practical in large-scale inventories.
- Double sampling for stratification simple and efficient way to utilize prior information.
- Method of Grafström & Tillé simple and apparently efficient way to balance 2nd-phase sample spatially.
- Very much work in progress; interesting to see
  - how **bal2** fares, when prior info not so good
  - effect of number of strata: effect of random sample size of sys2 pronounced with more strata
    $\Rightarrow$ smaller samples within some strata

Use prior information!

Thank you!

NATURAL RESOURCES
INSTITUTE FINLAND