Using auxiliary data to improve accuracy of NFI parameter estimates

A case study based on normalized digital surface model and NFI1 data

Radim Adolt

ÚHÚL Brandýs nad Labem, Czech Republic

ForestSAT 2014, 4.-7. November 2014

Contents

Availability of auxiliary data

Looking some years back, there's no doubt that:

- **the amount and availability of RS data increases**
- **temporal** and **spatial resolution** follows this trend
- **the number of products** - processed from this raw data increases (at least at the same rate)
- there is **large potential for new applications in forestry**
- **NFIs and operational forest inventories** will use and rely on the available RS-derived ancillary data, perhaps even more than now-days

Improving target parameter estimation by RS-derived ancillary data is among the most promising (and challenging) research areas.

Two ways towards enhanced NFI estimation

After the sample has been drawn

- **auxiliary data not used to modify sampling design** - sample locations are selected independently on the ancillary information
- **accuracy improvement** thanks to field and ancillary data correlation
- **design-based regression and difference estimators** - various kinds of these depending on the nature of ancillary data, includes the well-known **post-stratification**

Two ways towards enhanced NFI estimation

Before the sample has been drawn

- **auxiliary data influences sampling** - sample locations selected in a way, which depends on the auxiliary data
- **stratified sampling** - varying density of sample location in spatially defined strata
- **IDPI** - Inclusion Density Proportional to Intensity (PPI), a continuous-sampling analog to PPS (Probability Proportional to Size)
- **balanced sampling** - e.g. cube method [Deville and Tillé, 2004], defined and developed for finite populations, transferable to continuous sampling
- **accuracy improvement** either due to different (reasons or their combinations)
  - stratified sampling - optimized ratio of sample and stratum size (or/and within stratum variance)
  - IDPI - focusing samples to areas with higher amount of resource, total of which is to be estimated
  - balanced sampling - one phase estimator of auxiliary variable(s) equals to true total(s), analogy to **calibration property of GREG**
- parameters estimated usually by **one-phase, design-based** techniques

Pros and cons of GREG et al. I

Pros

- **sampling independent on the auxiliary data** - auxiliary data does not need to exist at the moment of sampling
- **new and better auxiliary data coming year by year**
- selection of most suitable data can be done just in the moment of estimation
- different auxiliary data can be used for different target parameters
- constant sampling density (expected number of sample locations per unit area) generates **representative samples** - the sample tends to be a minimized version of the whole population [Grafstrom and Schelin, 2013]
- representative samples (thanks to constant sampling intensity) are **best for permanent plots - long term monitoring**

Pros and cons of GREG et al. II

Additivity of total estimators

Additivity[a] **of total estimator** $\hat{Y}$ is defined by

$$\hat{Y}_{\bigcup_{i=1}^{k} D_i} = \sum_{i=1}^{k} \hat{Y}_{D_i}. \tag{1}$$

An additive total estimator evaluated for arbitrary domain being an union $\bigcup_{i=1}^{k} D_i$ of k sub-domains $D_1, D_2 \ldots D_k$ equals to the sum of k total estimators of the same kind calculated individually for each sub-domain.

[a] Geographical additivity or additivity defined by different non-spatial attributes of the population.

Pros and cons of GREG et al. III

Cons

- GREGs (Generalized ReGresion estimators) with linear models **parametrized for several target areas separately are <u>not additive</u>** - sums of estimates for divisions of the whole study area do not match the estimate for the whole
- SREG (Survey ReGression estimator) - solves the above problem, linear model is parametrized using data from the whole study area and applied to its divisions
- GREGs (SREGs) are **<u>not additive</u>, if different model formulations and/or auxiliary variables are used** for any partition of a target variable - e.g. GREGs for conifers + broadleaved $\neq$ GREG for all species
- GREGs have some **<u>limits in small areas defined as attribute-partitions</u>**
- if GREGs (SREGs) are applied with **one model formulation to several variables**, quite often, in some target variables or/and small geographic areas **<u>precision falls below the level of one-phase estimator</u>**
- GREGs **need enough samples (sample points) to parametrize** the model and to have the desired statistical properties

Cube and LPM fused [Grafstrom and Tille, 2012]

Properties of the new method

- combines **Cube method** by Deville and Tillé [2004] (balanced sampling, PPS) and the **Local Pivotal Method** (LPM, regular sampling in space) by Grafstrom et al. [2012]
- sampling method for **finite populations** of elements located in space (spatially correlated attributes)
- transferable to continuous NFI sampling as defined by Mandallaz [1991]
- **samples which are balanced** - definition in [Grafstrom and Tille, 2012]
- **samples which are well spread** over space - incorporates the LPM by [Grafstrom et al., 2012]
- **IDPI - PPS samples** - probability proportional to whatever (optimal)
- **LCUBE estimatrors are unbiased and additive** - no matter what the algorithm setup is

Evaluation of the new LCUBE method I

Study area(s)

- the whole Czech Republic (79 k. sq. km)
- **on this level linear model of the SREG was parametrized**
- **two target regions at the NUTS4 level**, where **we knew that SREG fails for the less frequent species**
  1. Břeclav (103798 ha) - **high proportion of broadleaved, few conifers**
  2. Žďár nad Sázavou (157855 ha)- the ratio switched

Evaluation of the new LCUBE method II

The population and auxiliary data

- population of **per ha growing stock for all, coniferous and broadleaved species** - attribute domains
- source SFMPs (**Summarized Forest Management Plans**), time window 2012-1994, coverage of more than 93% (2.6 mill. ha) of forest land within Czech Republic (2.8 mill. ha)
- **auxiliary data - nDSM originated by SRTM** (2001, NASA) minus **DTM4g** (ČÚZK, 2010-2013, 5m resolution)
- both data sets in a form of a raster with **250m pixel-size**
- linear model used by SREG explains **50.2% of the overall variance**[a] **of all species, 43.9% coniferous and 9% of broadleaved**

[a] Total variance measured over the whole country including non-forest land with 0 growing stock.

Evaluation of the new LCUBE method III

The designs and estimation - CSS

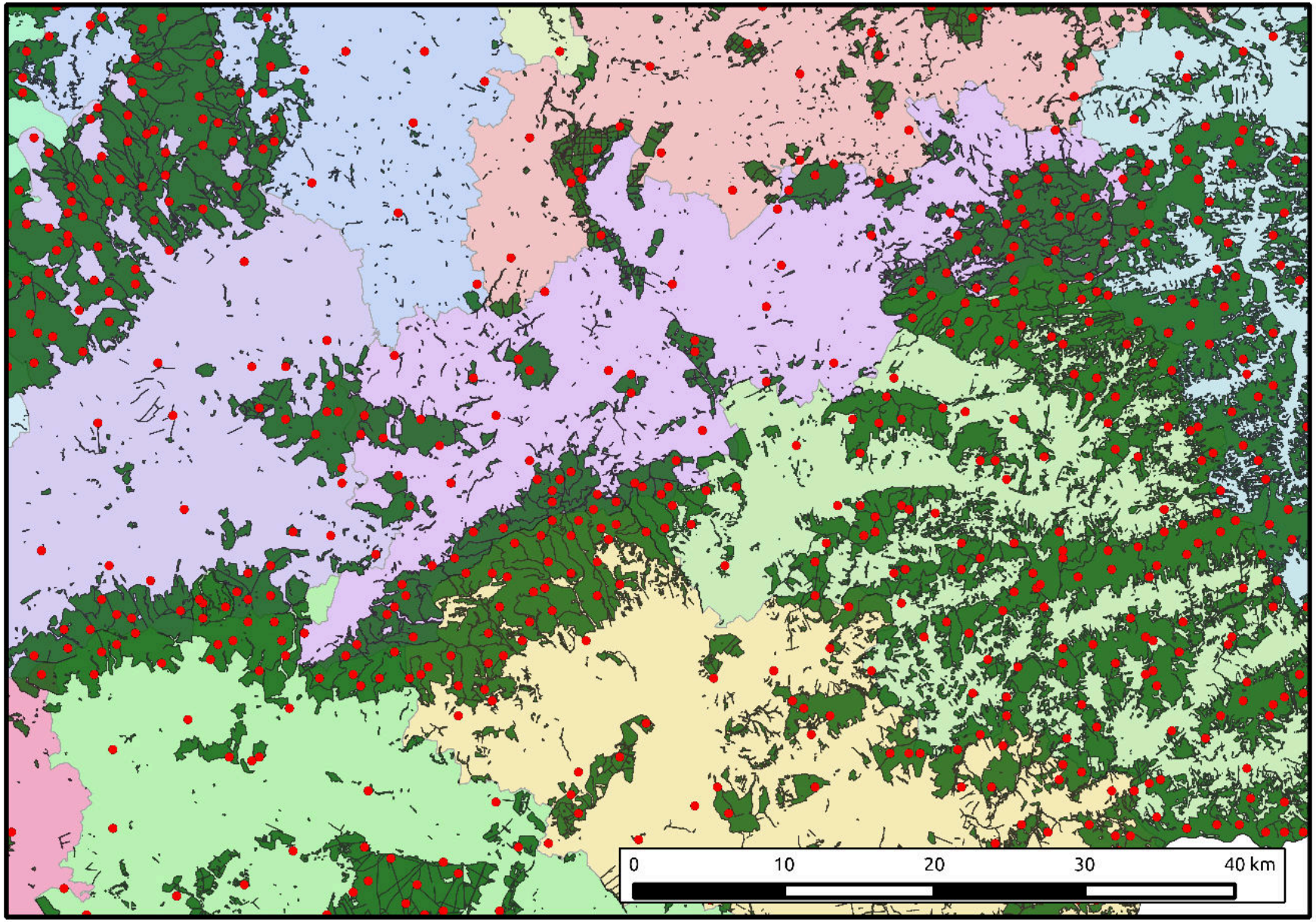
- **Aligned (or centric) Systematic Sampling (CSS) with 3km block size** (approx. 9000 sample locations within Czech Republic in each simulation) and random origin was repeated 300 times
- population and auxiliary values were taken on sample points intersecting the two raster layers (SFMP population and nDSM auxiliary data)
- **one-phase and two-phase totals** (SREG with linear model parametrize on the level of the whole country) were estimated for each replication of the design, all species, coniferous and broadleaved separately, and both of the NUTS4 regions

Evaluation of the new LCUBE method IV

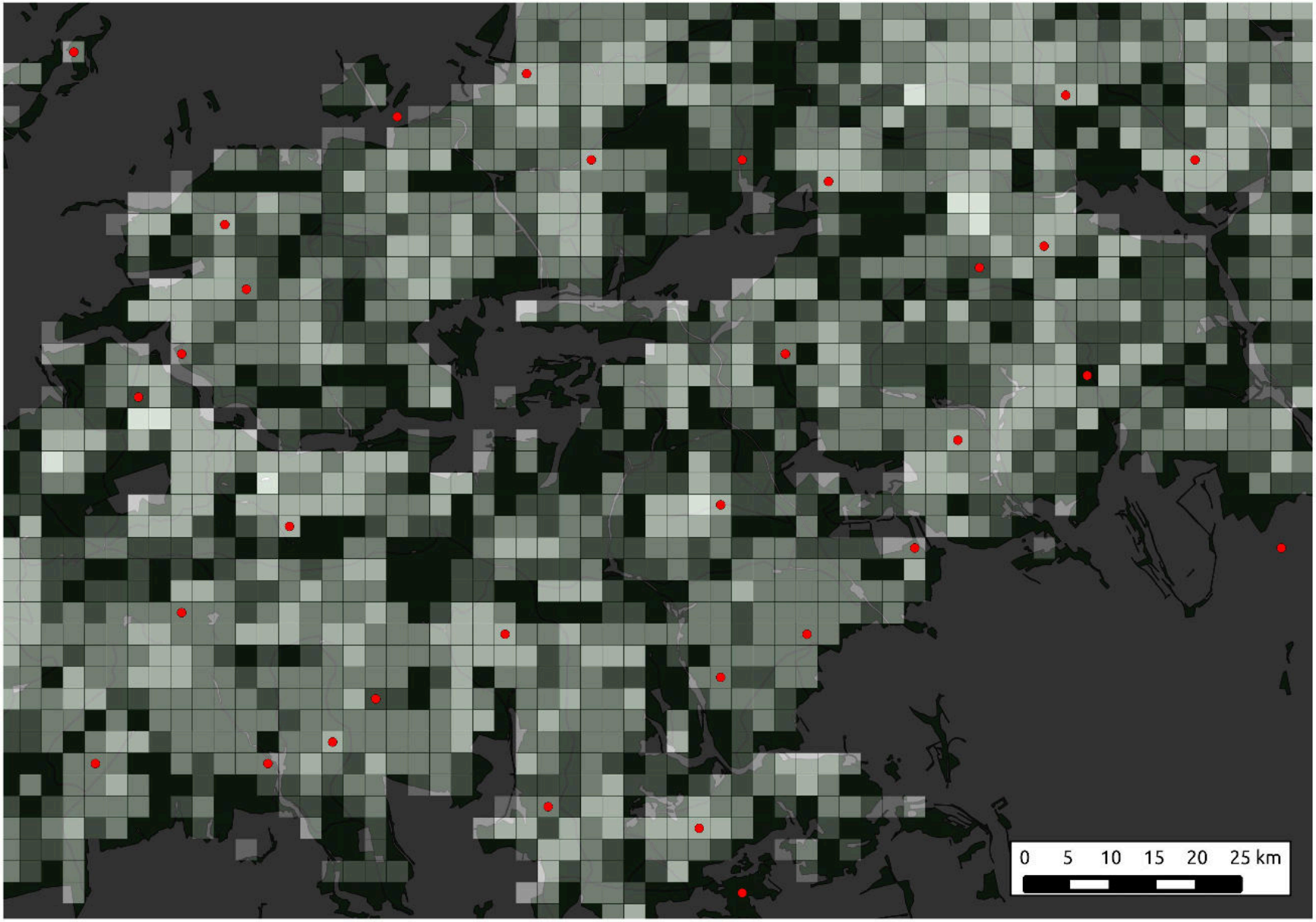
The designs and estimation - LCUBE

- 300 replications of the alternative design were generated using LCUBE method [Grafstrom and Tille, 2012], implemented in R (**package BalancedSampling**)
- **balancing defined** on the nDSM, x-coordinate, y-coordinate, the square of the two
- **LPM part of the algorithm** used x- and y-coordinate as the only spread variables
- **unequal probabilities** were calculated as the ratio of individual pixel values to the total of nDSM within the whole Czech Republic multiplied by the required expected number of sample locations in the whole country (9000)
- for the purpose of probability calculation, all zero pixels with zero nDSM value got new value equal to one
- the algorithms were run separately in each of the 77 NUTS4 units
- after intersections with population and auxiliary layers values of **one-phase estimators** were calculated taking unequal inclusion densities into account - see Cordy [1993]

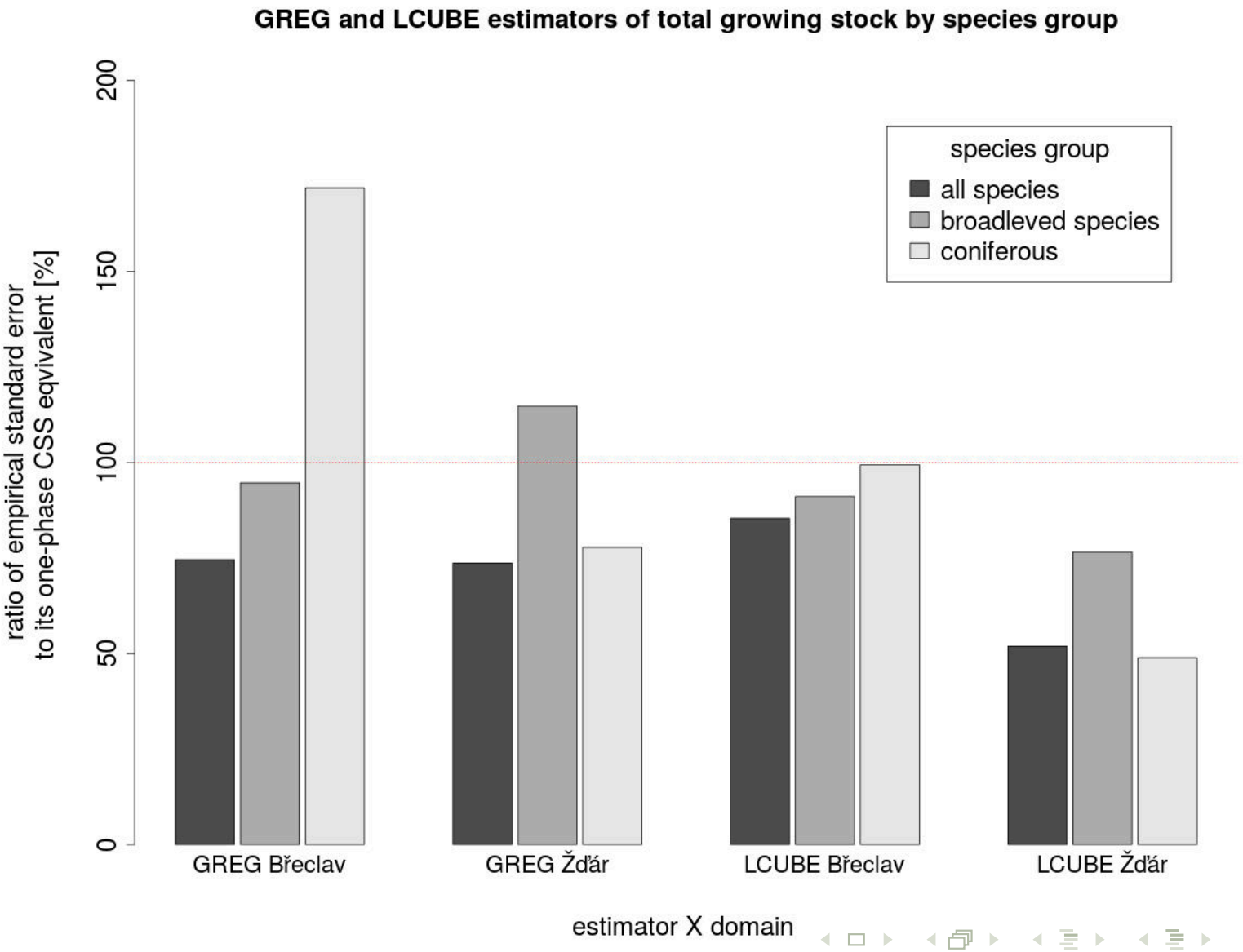
LCUBE sample



LCUBE sample



Results


GREG and LCUBE estimators of total growing stock by species group
ratio of empirical standard error
to its one-phase CSS eqvivalent [%]
0
50
100
150
200
species group
all species
broadleved species
coniferous
GREG Břeclav
GREG Žďár
LCUBE Břeclav
LCUBE Žďár
estimator X domain

Lessons learned

- in general **LCUBE performed better** on small domains than GREG (given equal overall sample sizes)
- this result seems to be mainly **because of the IDPI usage**, balancing may help to a degree comparable to GREG (or even more?)
- using auxiliary data of this quality (SRTM, reduced spatial resolution, $R^2 = 50\%$) **LCUBE has not delivered better estimates, if the number of sample LCUBE locations was reduced to the amount actually surveyed in the field** by the CSS design [1]
- if the relative performance of GREG and LCUBE remains unchanged for auxiliary information **with elevated R^2, LCUBE might be competitive** in practical NFI situations (skipping measurements on non-forest plots)
- **generation of LCUBE samples takes a lot of time**

[1] In many NFIs clearly non-forest plots are not assessed in the field.

Conclusion (partial) and further directions I

IDPI implemented by PPS-LCUBE - non-representative samples

- from the perspective of NFIs and even more operational forest inventory[a], **LCUBE by [Grafstrom and Tille, 2012] is worth further attention**
- **when combined with IDPI representativeness gets lost** - for variables not used to define inclusion probabilities
- as the population changes in time, IDPI sample looses its effect - **not suitable for long-term monitoring on permanent plots** as the only sampling approach
- evaluation of the accuracy gain should **take into account the number of plots actually surveyed in the field** for the alternative design (NFI specificity)
- to get an advantage of PPS-LCUBE[b] **auxiliary information must explain quite a high proportion of the overall variance**, depending on the share of plots excluded from the field survey during NFIs

[a] An exact delineation of forest area to survey.

[b] As well as other IDPI alternatives e.g. the GRTS by Stevens and Olsen [2004] combined with PPS

Conclusion (partial) and further directions II

Non IDPI, non-PPS LCUBE - representative samples

- **non-IDPI (non-PPS) LCUBE produces well spread representative samples** balanced on preselected set of variables - this setting has not the limitations of IDPI mentioned before, balancing can **use variables which are stable in time**
- **not much empirical evidence** on how non-PPS LCUBE works in comparison to conventional (CSS or similar) NFI sampling followed by GREGs (SREGs)
- in theory, balancing should work well even less frequent population fractions, because **LCUBE uses the whole auxiliary population to balance samples**
- deriving g-weights **GREGs work only with pairs of ground-truth/auxiliary data** - problems with parametrization of models specific for very small domains (in the geographic as well as attribute sense)
- **in very small domains, non-PPS LCUBE could work even better than GREGs**

End of presentation I

Thank You For Attention!

Literature I

C. B. Cordy. An extension of the horwitz-thompson theorem to point sampling from a continuous universe. *Statistics and Probability Letters*, 18:353–362, 1993.

J-C Deville and Y. Tillé. Eficient balanced sampling: The cube method. *Biometrika*, 91:893–912, 2004.

A Grafstrom and L. Schelin. How to select representative samples. *Scandinavian Journal of Statistics*, 41:277–290, 2013.

A Grafstrom and Yves Tille. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24: 120–131, 2012.

A Grafstrom, P. Lundstom, L., and L. Schelin. Spatially balanced sampling through the local pivotal method. *Biometrics*, 2012.

Literature II

D. Mandallaz. *An unified approach to sampling theory for forest inventory based on infinite population and superpopulation models*. PhD thesis, Swiss Federal Institute of Technology (ETH), Zurich, 1991.

D. L. Jr. Stevens and A. R. Olsen. Spatialy balanced sampling of natural resources. *Journal of the American Statistical Association*, 99:262–278, 2004.