Domain estimation and poststratification in national forest inventory

Adrian Lanz[1] Radim Adolt[2] Jiří Fejfar[2] Berthold Traub[1]

[1] Swiss Federal Institute for Forest, Snow and Landscape Research WSL

[2] Forest Management Institute (FMI), Branch Kroměříž

2nd International Workshop on Forest Inventory Statistics,
24th - 26th May 2016, Kroměříž, Czech Republic

Contents

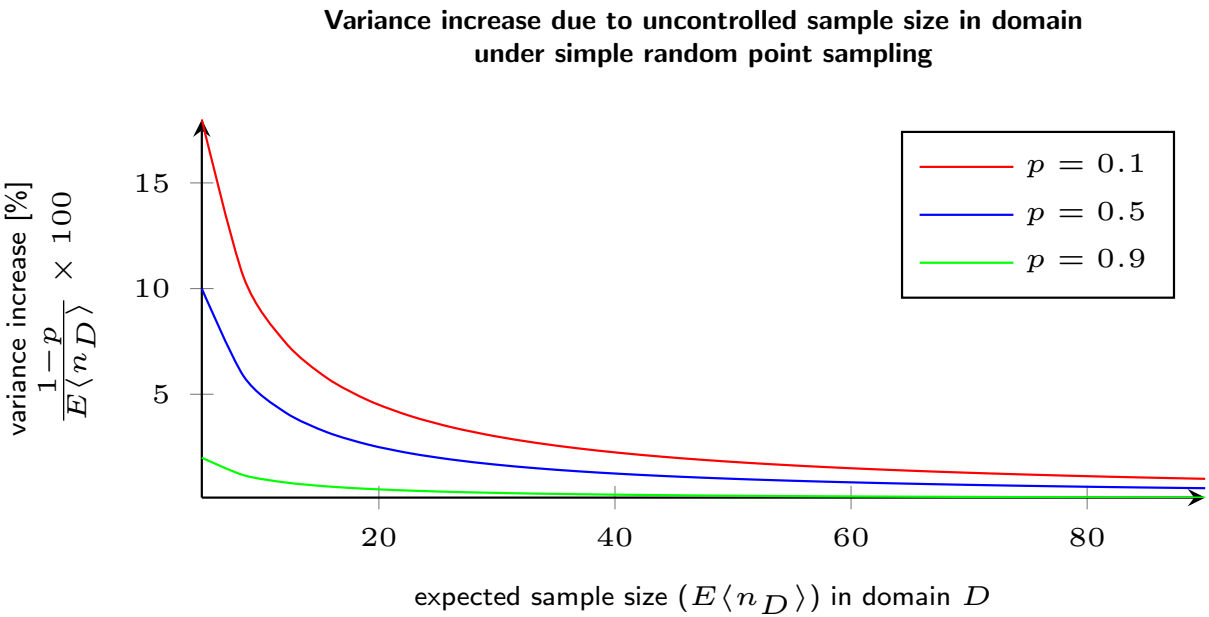
Problem statement

- we assume a (national) forest inventory with an uniform distribution of sample plots over a (sampling frame of) known surface area
  - typically a sampling grid (systematic sampling)
  - sampling frame may cover the entire country, or be partitioned into disjoint sampling strata with independent designs
- and the need for estimates in sub-regions, so-called domains (of interest)
  - NUTS regions (domains of known surface area)
  - forest types (domains of unknown surface area)
- topic of this study: options and statistical properties of domain estimation

Domain with known size (I)

Estimators and theoretical variance of estimates

- the sample size n_D in domain D is random
- the estimator for the mean ($\hat{Y}_D = \frac{\sum_{j \in D} y_j}{\sum_{j \in D} \iota_{D.j}}$) and the total ($\hat{T}_D = \lambda_D \hat{Y}_D$) are (quasi) design-unbiased
- but the true unconditional variance of the estimates is higher than the variance under controlled sample size
- effect depends on the expected sample size n_D and on the relative size of the domain $p_D = \frac{\lambda_D}{\lambda_S}$ (λ_S is the size of the sampling frame)



Subdomain of known size (II)

Variance estimation

- in estimation, we ignore the fact that the sample size is random and condition the variance estimate on the realised size of the sample
- the unconditional (true) variance remains of interest in sampling design planning
- with $\hat{S}_D = \frac{\sum_{j \in D}(y_j - \bar{y}_D)^2}{n_d - 1}$, the empirical variance of the target variable y in domain D, approximately design-unbiased estimators of the variance of $\hat{Y}_D$ under simple random point sampling and when conditioning on the realised sample size n_D are

$$\hat{V}\langle \hat{Y}_D \,|n_D\rangle = \frac{\hat{S}_D^2}{n_D} \qquad \text{standard conditional}$$

$$\hat{V}\langle \hat{Y}_D \,|n_D\rangle = \frac{\hat{S}_D^2}{n_D} \frac{\mathring{n}_S(n_D - 1)}{n_D(\mathring{n}_S - 1)} \qquad \text{standard ratio (Taylor linearisation)}$$

$$\hat{V}\langle \hat{Y}_D \,|n_D\rangle = \frac{\hat{S}_D^2}{n_D} \frac{(n_D - 1)}{n_D} \qquad \text{model-assisted (g-weights)}$$

- effect: we get, by chance and depending on the realised sample size, a relatively high or low precision of the domain estimates

Domain of unknown size (I)

Point estimates

- mean estimator is the same as with known surface area
- total estimator can be understood and written in two forms

$$\hat{T}_D = \frac{1}{\mathring{n}_S} \sum_{i=1}^{\mathring{n}_S} \iota_{D.j} y_j = \frac{1}{\mathring{n}_S} \sum_{i=1}^{\mathring{n}_S} y_{D.j} \qquad \text{standard: with domain indicator}$$

$$= \sum_{j \in D} \frac{\lambda_S}{\mathring{n}_S} y_j \qquad \text{weighted sum over plots in } D$$

- the total estimator is unconditionally, over repeated samples, design-unbiased, but may be conditionally, for the realised sample, biased with

$$B\langle \hat{T}_D \,|n_D\,\rangle = \left(\frac{\lambda_S \, n_D}{\mathring{n}_S \, \lambda_D} - 1 \right)$$

- which is zero, if the realised sample size is, by chance, such that $\frac{n_D}{\mathring{n}_S} = \frac{\lambda_D}{\lambda_S}$

Domain of unknown size (II)

True variance

- conditioning on the realised sample size, we get for this conditional mean squared error of $\hat{T}_D$

$$MSE\langle \hat{T}_D \,|n_D\rangle = (B\langle \hat{T}_D \,|n_D\rangle)^2 + V\langle \hat{T}_D \,|n_D\rangle$$
$$= \lambda_D^2 \left(Y_D^2 \left(\frac{\lambda_S n_D}{\mathring{n}_S \lambda_D} - 1 \right)^2 + \frac{S_D^2}{n_D} \left(\frac{\lambda_S n_D}{\mathring{n}_S \lambda_D} \right)^2 \right)$$

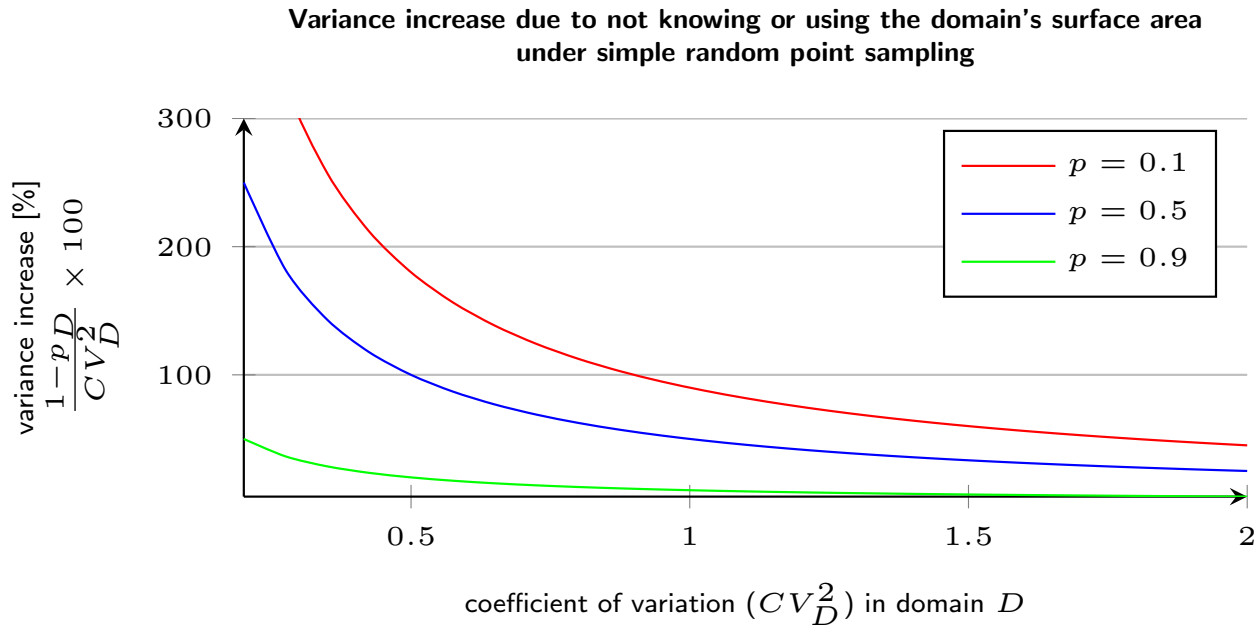
- conditioning on samples of size is $n_D = \mathring{n}_S \frac{\lambda_D}{\lambda_S}$, the bias is zero and the mean squared error is equal to the variance of $\hat{T}_D$ under a fixed sample size n_D and a known domain size λ_D
- unconditionally, however, the true variance is higher and the relative efficiency compared to estimating with known surface area is

$$\rho = \frac{V\langle \hat{T}_D \rangle_{\lambda_D \text{ unknown}}}{V\langle \hat{T}_D \rangle_{\lambda_D \text{ known}}} = 1 + \frac{1 - p_D}{CV_D^2}$$

- thus, the loss due to not using or having available the true surface area is higher for small domains ($p_D \to 0$) and for homogeneous domains ($CV_D^2 = \frac{S_D^2}{Y_D^2} \to 0$)

Domain of unknown size (III)

Loss in precision illustrated



- typical coefficients of variation (CV_D^2) from the Swiss NFI
  - growing stock on forest land: 0.5 to 0.6
  - gross increment on forest land: 0.7 to 0.8
  - growing stock *Quercus robur* on forest land: 16 to 484 (CH: 57)
  - growing stock on total land: 3 to 6 (with a proportion of forest land between 40% and 20%)
  - area estimation: $CV_D^2 = 0$!

Domain of unknown size (IV)

Intermediate conclusions

When estimating totals for (geographic) domains of unknown size:

- try to find a frame $S \supseteq D$ of known size λ_S, such that the relative size of the domain (p_D) is as close to 1 as possible
- for a target variable with a high coefficient of variation in the domain ($CV_D^2 \to \infty$): the loss in precision due to not knowing the domain size is possibly small
- in large-area (national) forest inventory, the forest land itself is usually a domain of unknown size

With an external estimate of the domain's surface area

- the total estimator is $\hat{T}_D = \hat{\lambda}_D \, \hat{Y}_D$
- we assume independent and probability sampling based estimates $\hat{\lambda}_D$ and $V\langle\hat{\lambda}_D\rangle$
- using a general result from (Goodman 1960) about the variance of the product of two independent variables, we get

$$V\langle\hat{T}_D\rangle = V\langle\hat{\lambda}_D\rangle \, (E\langle\hat{Y}_D\rangle)^2 + V\langle\hat{Y}_D\rangle \, (E\langle\hat{\lambda}_D\rangle)^2 + V\langle\hat{\lambda}_D\rangle \, V\langle\hat{Y}_D\rangle$$

- and the proposed estimator is

$$\hat{V}\langle\hat{T}_D\rangle = \hat{V}\langle\hat{\lambda}_D\rangle \, \hat{Y}_D^2 + \hat{V}\langle\hat{Y}_D\rangle \, \hat{\lambda}_D^2 - \hat{V}\langle\hat{\lambda}_D\rangle \, \hat{V}\langle\hat{Y}_D\rangle$$

- yes, the minus sign for the last term in the estimator is correct and not a typing error!

The random number of points in domain under simple random point sampling is not an efficient estimate of the domain size!

- we may be tented to use the relative number points in domain as an estimate of the domain's surface area
- but under simple random point sampling of $\mathring{n}_S$ points in frame S of size λ_S we find

$$\hat{T}_D = \hat{\lambda}_D \hat{Y}_D$$

$$\hat{\lambda}_D = \frac{\lambda_S}{\mathring{n}_S} n_D$$

$$V\langle \hat{\lambda}_D \rangle = (\frac{\lambda_S}{\mathring{n}_S})^2 V\langle n_D \rangle$$

$$V\langle n_D \rangle = \mathring{n}_S \, p_D (1 - p_D)$$

- and the algebraic transformation confirms that the variance of total estimate is not improved in this case (which is intuitively understandable as this estimator does not rely on additional information in the estimation of the domain's size)

Two phase sampling (I)

Loss in precision due to now using/knowing the domain size, but with a first phase sample of points indicating the domain without error

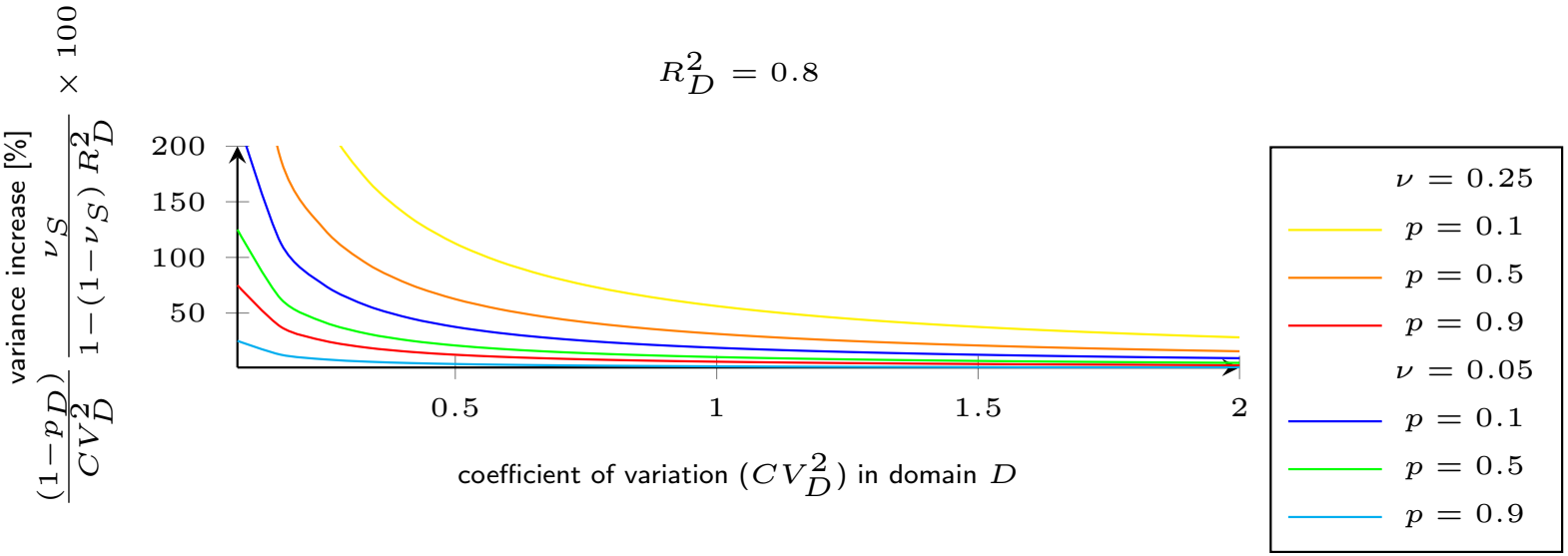
- second phase sample selection: simple random sampling without replacement
- the relative efficiency under two phase sampling compared to estimation with known domain size is (Mandallaz, internal report)

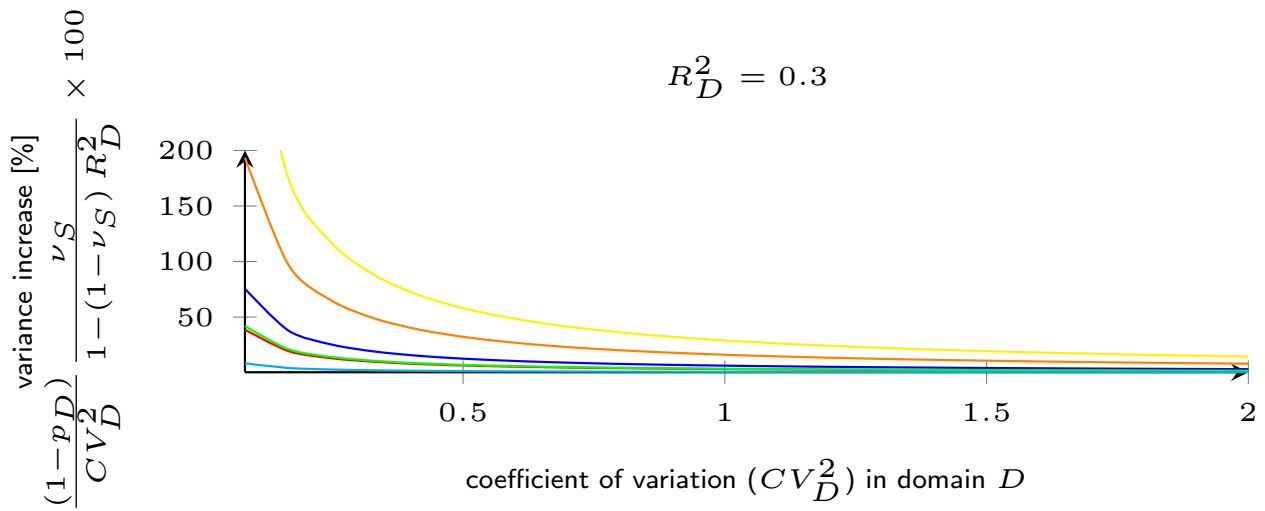
$$\rho = \frac{V\langle \hat{T}_{D.\text{2phase}} \rangle_{\lambda_D \text{ unknown}}}{V\langle \hat{T}_D \rangle_{\lambda_D \text{ known}}} = 1 + \frac{(1-p_D)}{CV_D^2} \frac{\nu_S}{1-(1-\nu_S)\,R_D^2}$$

- R_D^2, coefficient of determination of the model linking the target variable y in D with auxiliary information available on first phase sampling plots
- $\nu_S = \frac{\mathring{n}_{S_2}}{\mathring{n}_{S_1}}$, the fixed proportion of second phase sample plots in frame

Two phase sampling (II)

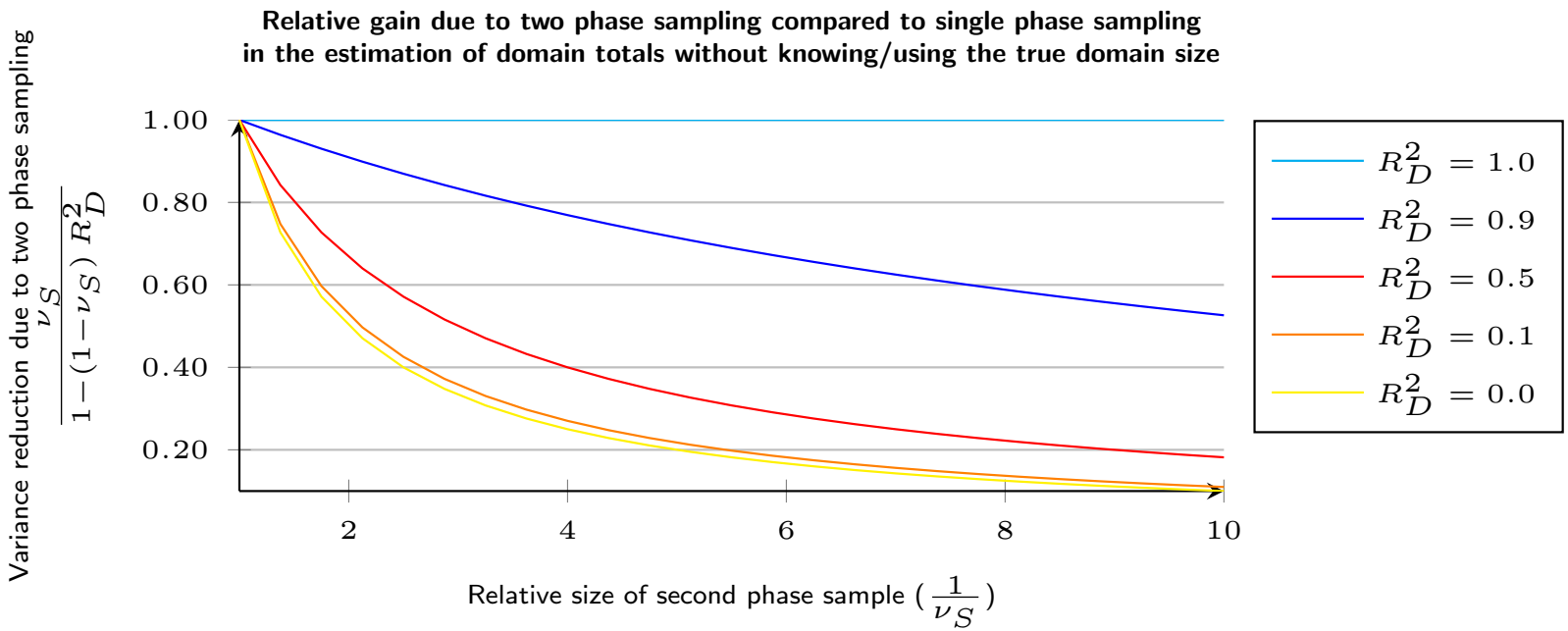
The variance increase due to knowing/using the domain's surface area depends on the coefficient of determination of the model (R^2_D) and the coefficient of variation of the target variable (CV^2_D) in the domain, as well as on the relative size of the domain (p_D) and the second phase sample (ν).





Two phase sampling (III)

- the better the model in domain D ($R^2_D \to 1$), the higher counts the loss due to not having the domain's true size available
- with a minor model ($R^2_D \to 0$): the first phase sample points indicating the domain allow for great relative gains
- the general recommendation from single phase sampling remains: find a frame with $p_D \to 1$



Systematic sampling (I)

Domain size known

- the mean is (quasi) design-unbiased, although the sample size n_D in D is random
- the (quasi) design-unbiased estimator fot the total in domain D under systematic sampling is therefore

$$\hat{T}_D = \lambda_D \hat{Y}_D = \lambda_D \frac{\sum_{j \in D} y_j}{\sum_{j \in D} \iota_{D.j}} = \lambda_D \bar{y}_D$$

- if the distribution of the target variable y over the domain is of no particular order with respect to the variance of the mean estimator, we can use the same conditional variance estimator as under simple random point sampling

$$\hat{V}\langle \hat{T}_D \,|n_D\rangle = \lambda_D^2 \hat{V}\langle \hat{Y}_D \,|n_D\rangle = \lambda_D^2 \frac{\hat{S}_D^2}{n_D} = \lambda_D^2 \frac{\sum_{j \in D}(y_j - \bar{y}_D)^2}{n_D(1 - n_D)}$$

which estimates the true variance over repeated samples of size n_D

Systematic sampling (II)

Domain size unknown

- analogy to simple random point sampling with the area represented by point $\frac{\lambda_S}{\mathring{n}_S}$ replaced by the grid cell size λ_C
  - the total estimator

$$\hat{T}_D = \lambda_C \sum_{j \in D} y_j$$

    is exactly design-unbiased, but has conditional bias of

$$B\langle \hat{T}_D \,|n_D\,\rangle = T_D(\frac{n_D\,\lambda_C}{\lambda_D} - 1)$$

  - the true (unconditional) variance under repeated sampling and assuming that the distribution of the target variable y over the domain is of no particular order with respect to the variance of the mean estimator, is approximately

$$V\langle \hat{T}_D \rangle = \lambda_C^2 \left(Y_D^2\, V\langle n_D \rangle + S_D^2\, E\langle n_D \rangle + \frac{V\langle n_D \rangle}{E\langle n_D \rangle} S_D^2 \right)$$

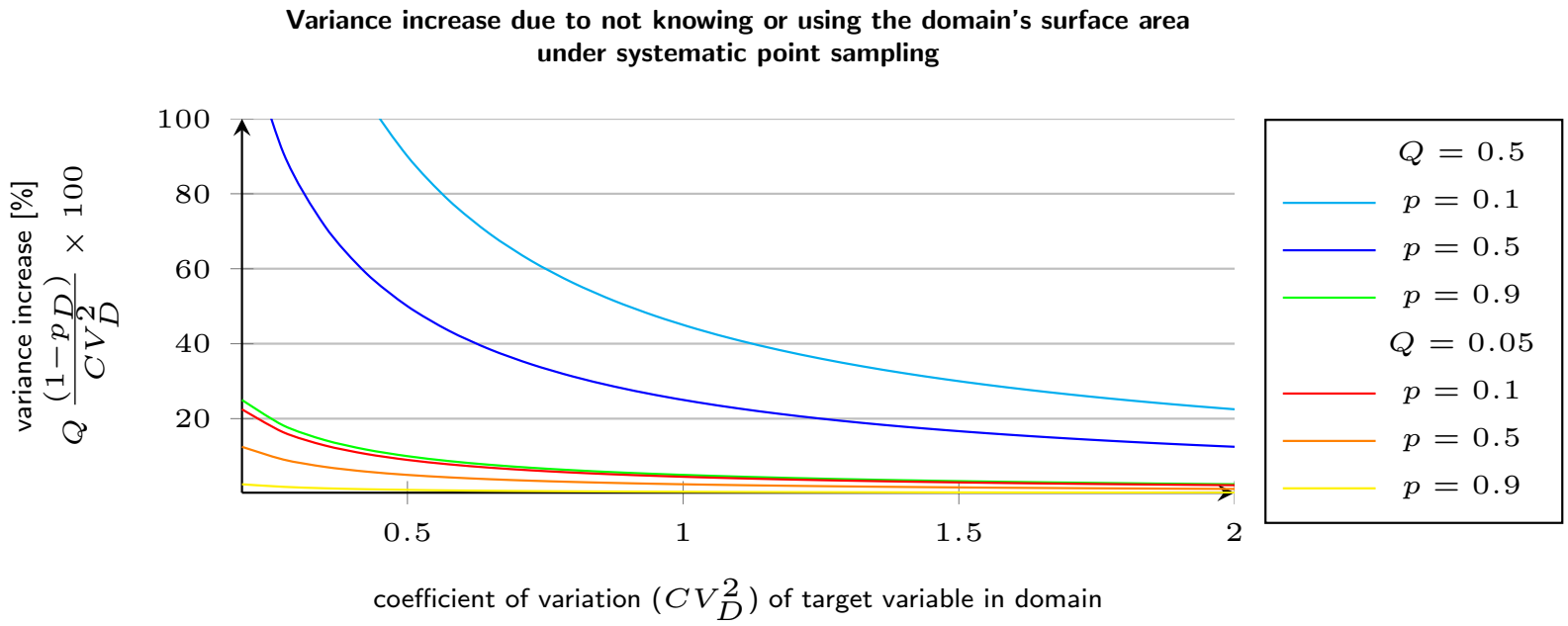
- under systematic sampling, $V\langle n_D \rangle$ can be expected to be smaller than under simple random point sampling

Systematic sampling (III)

Domain size unknown

- defining $Q = \frac{V\langle n_D \rangle_{\text{sys}}}{V\langle n_D \rangle_{\text{srs}}}$ and assuming that the distribution of the target variable y over the domain is of no particular order with respect to the variance of the mean estimator, we get a relative efficiency of the domain total estimator under systematic sampling compared to the total estimator with known domain size of

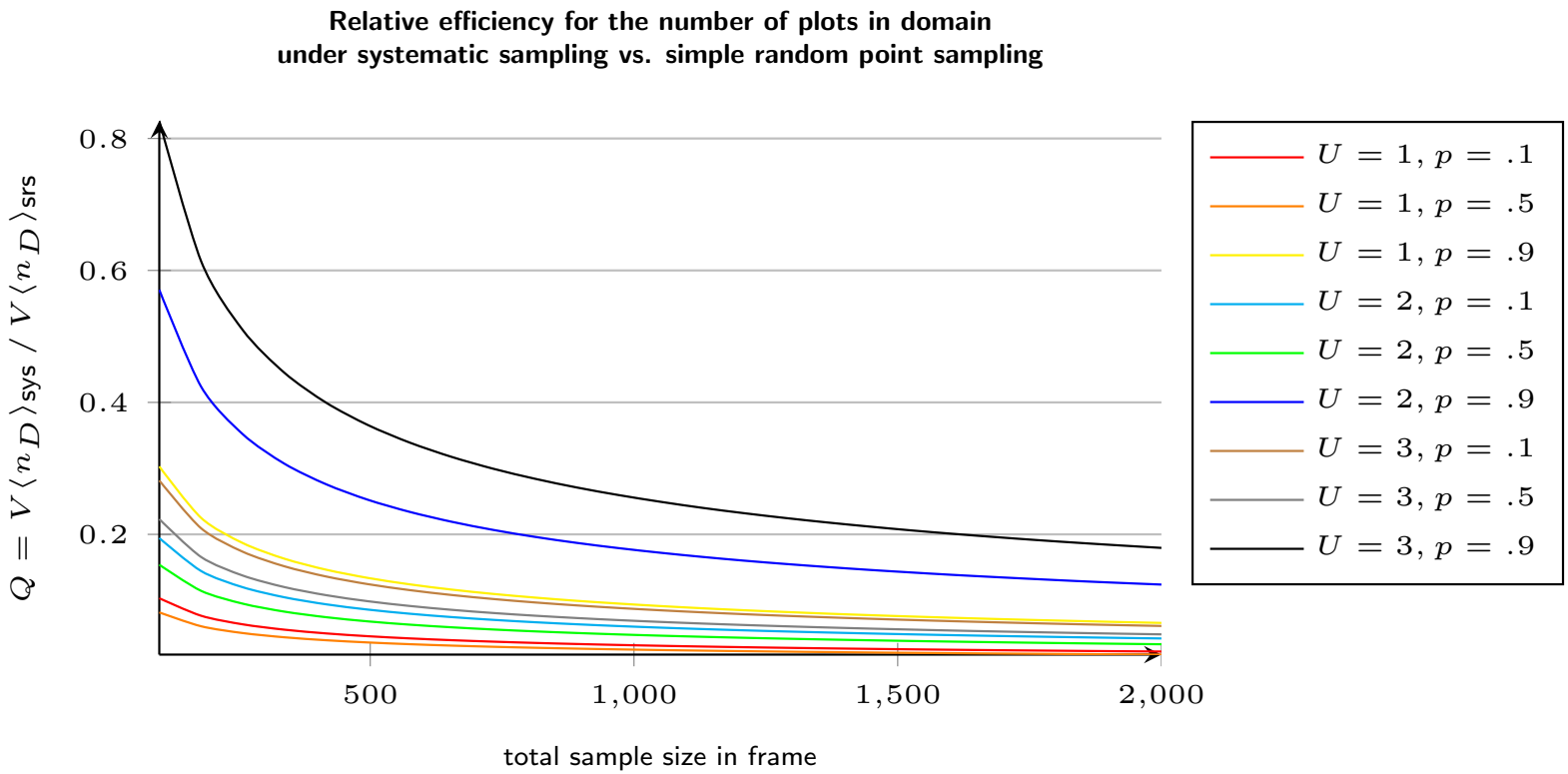
$$\rho = \frac{V\langle \hat{T}_D \rangle_{\lambda_D \text{ unknown}}}{V\langle \hat{T}_D \rangle_{\lambda_D \text{ known}}} = 1 + Q\,\frac{(1-p_D)}{CV_D^2}$$



Systematic sampling (IV)

How small is $Q = \frac{V\langle n_D\rangle_{\text{sys}}}{V\langle n_D\rangle_{\text{srs}}}$ in realistic cases? A model example...

- Zöhrer/Kleinn: $V\langle n_D\rangle \approx (E\langle n_D\rangle)^2\,(10^{\log(S_{\%})})^2\,100^{-2}$
- with $\log(S_{\%}) = 1.739 - 0.755\log(n_D) + 0.457\log(U_{D.r})$
- and $U_{D.r}$, the ratio between the true perimeter of the domain divided by the circumference of the circle of the same surface area



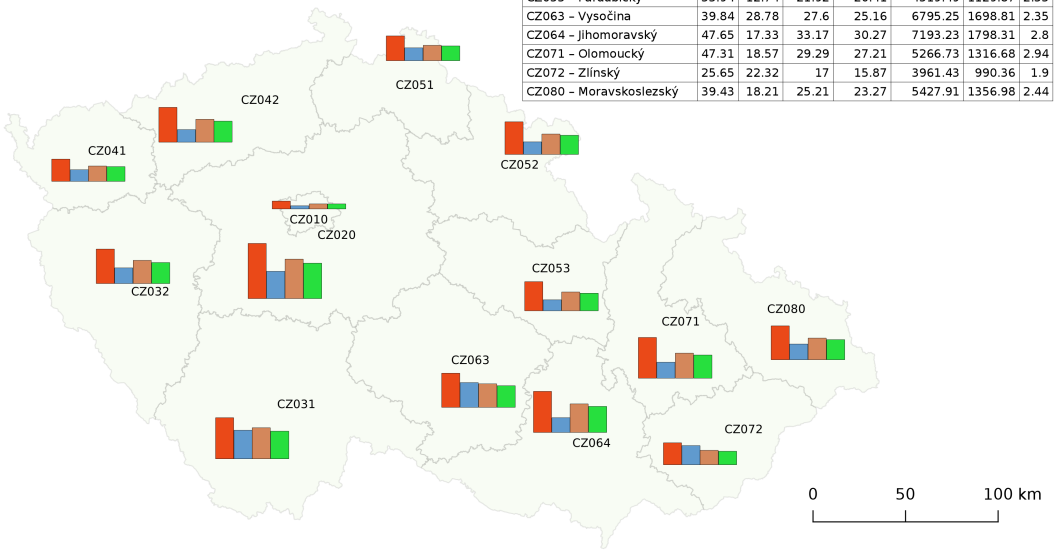
Systematic sampling (V)

Some $Q = \frac{V\langle n_D \rangle_{\text{sys}}}{V\langle n_D \rangle_{\text{srs}}}$ calculated for NFI polygons with Zöhrer's formula

- 26 cantons of Switzerland
  - 0.004 to 0.010 (8 points per 2 sqkm)
  - 0.010 to 0.070 (1 point per 2 sqkm)
  - 0.040 to 0.200 (1 point per 18 sqkm)
- 5 production zones of Switzerland
  - 0.004 to 0.007 (8 points per 2 sqkm)
  - 0.013 to 0.022 (1 point per 2 sqkm)
  - 0.400 to 0.700 (1 point per 18 sqkm)
- NUTS3 regions in the Czech Republic
  - 0.015 to 0.050 (1 point per 2 sqkm)
- although the perimeter is a difficult parameter in this model (fractional dimension), the results seem quite plausible (based on realised sample sizes and predicted $V\langle n_D \rangle$ in the above examples)

Systematic sampling (VI)

Comparing the Zöhrer/Kleinn model with simulated variances in NUTS3 regions of the Czech Republic



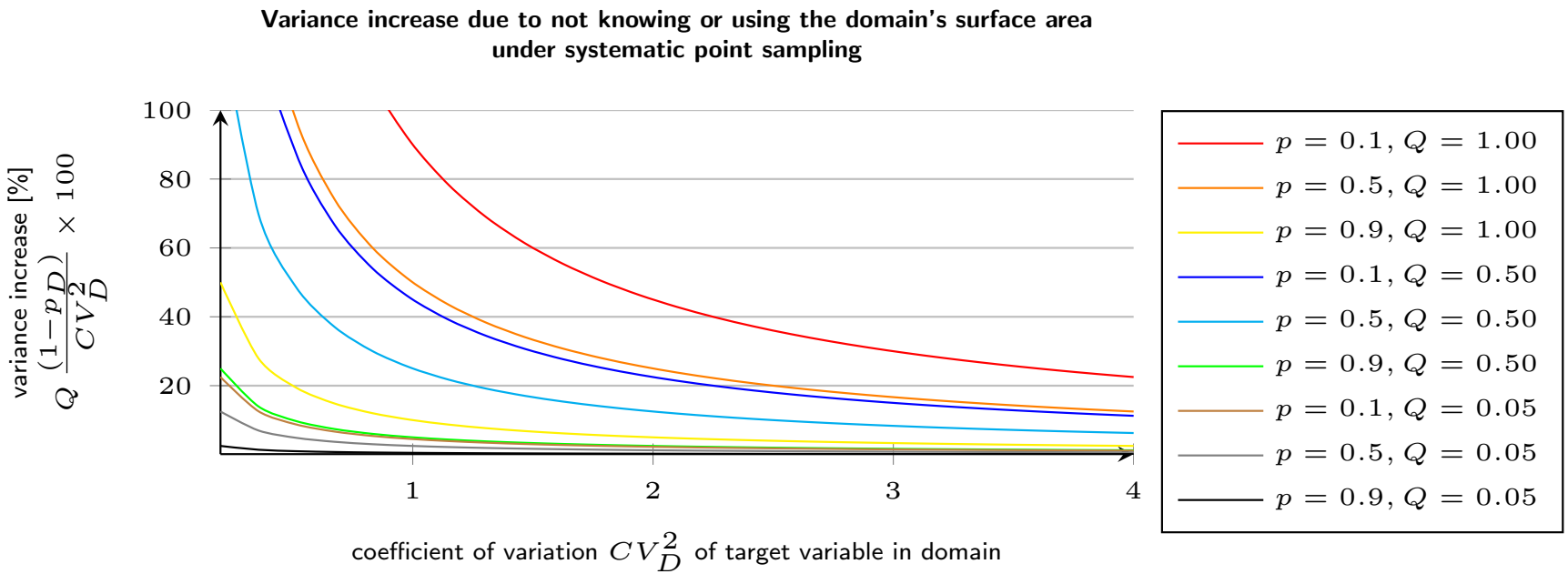
NUTS 3	NFI2	SYS S.	KLEINN	ZOHRER	area [km2]	exp_nd	UDr
CZ010 – Hlavní město Praha	9.38	3.93	5.88	6.13	496.18	124.04	2.04
CZ020 – Středočeský	64.18	31.68	46.06	41.2	11016.26	2754.06	3.12
CZ031 – Jihočeský	47.8	33.18	35.99	32.19	10058.06	2514.51	2.5
CZ032 – Plzeňský	40.04	18.3	27.07	24.49	7560.77	1890.19	2.16
CZ041 – Karlovarský	25.93	13.75	18.1	17.12	3313.94	828.48	2.27
CZ042 – Ústecký	40.59	14.81	26.55	24.58	5334.85	1333.71	2.61
CZ051 – Liberecký	28.76	14.99	18.09	17.16	3163.59	790.9	2.33
CZ052 – Královéhradecký	37.86	14.63	23.87	22.21	4759.15	1189.79	2.49
CZ053 – Pardubický	33.94	12.74	21.92	20.41	4519.49	1129.87	2.33
CZ063 – Vysočina	39.84	28.78	27.6	25.16	6795.25	1698.81	2.35
CZ064 – Jihomoravský	47.65	17.33	33.17	30.27	7193.23	1798.31	2.8
CZ071 – Olomoucký	47.31	18.57	29.29	27.21	5266.73	1316.68	2.94
CZ072 – Zlínský	25.65	22.32	17	15.87	3961.43	990.36	1.9
CZ080 – Moravskoslezský	39.43	18.21	25.21	23.27	5427.91	1356.98	2.44

- SYS2 are the true (simulated) variances (2 km by 2 km grid), however under fixed orientation of the grid
- the deviation between the empirical variance under the simulations and the predicted variance with the models from Zöhrer/Kleinn is relatively high in some NUTS3 regions (further tests needed)

Systematic sampling (VII)

Preliminary conclusions

- again, try to avoid small domains and small sample sizes
- the loss due to not knowing or using the surface area of the domain is relatively small for heterogeneous target variables (CV_D^2 large)
- the loss is larger for compact domains (high grid density, Q large)
- thus, the increase in variance due to no knowing or using the domain's surface area can be largely decreased if a good prediction or the true variance $V\langle n_D \rangle$ is available



The link to poststratification (I)

Some terminology

- domain of interest of known size
  - usually exists because of user information needs for the domain
  - or may otherwise be needed for estimation purposes
  - comprises a sub-region or the entire sampling space (sampling frame)
- post-sampling strata
  - for error reduction, possibly not user/stakeholder visible
  - a set of several disjoint sub-areas covering together the entire domain of interest
  - the size of the post-strata as well as the attribution of sample points to the post-strata must be known (both without error)
  - two phase sampling: post-sampling strata sizes may be estimated
- domain of interest of unknown size (attribute domain)
  - exists because of user information needs for the domain
  - a surrounding domain (or several disjoint domains) of known size is needed for estimation (and the attribution of sample points to the domain)
- subpopulation
  - preferred to be used for a subdomain of interest at tree level
  - sampling literature: domain and subpopulation are mostly synonyms

The link to poststratification (II)

Additivity of estimates for disjoint geographic domains

- we have explained, that the most precise total estimator for a domain e of known size λ_e

$$\hat{T}_e = \lambda_e \hat{Y}_e$$

- and the natural estimator for the overall total in a super-domain D consisting of E disjoint domains is then ($\lambda_D = \sum_{e=1}^{E} \lambda_e$)

$$\hat{T}_D^{(E)} = \sum_{e=1}^{E} \hat{T}_e$$

  which is the poststratified estimator with domains as strata
- but a second set of domains of known size may exist within the super-domain D ($\lambda_D = \sum_{f=1}^{F} \lambda_f = \sum_{e=1}^{E} \lambda_e$) and we get

$$\hat{T}_D^{(E)} = \sum_{e=1}^{E} \hat{T}_e \neq \sum_{f=1}^{F} \hat{T}_f = \hat{T}_D^{(F)}$$

- both estimators $\hat{T}_D^{(E)}$ and $\hat{T}_D^{(F)}$ are design-unbiased under repeated sampling, but for a given sample, the estimates are numerically different

Total estimates with two different sets of domains

Numerical difference in the estimates of above-ground biomass of live trees in Switzerland: example with production regions and biogeographical regions as post-sampling strata (the sum of domain totals equals the respective overall total, but the oveall totals are not the same)

NFI4b

biomass above ground of live trees : conifers/broadleaves · production region
unit: Mio kg
unit of evaluation: accessible forest without shrub forest
grid: grid NFI4 2009-2013
state 2009/13

	production region											
	Jura		Plateau		Pre-Alps		Alps		Southern Alps		overall	
conifers/broadleaves	Mio kg	± %	Mio kg	± %	Mio kg	± %	Mio kg	± %	Mio kg	± %	Mio kg	± %
confers	18715	5	23182	5	35027	4	48395	3	9688	8	135008	2
broadleaves	22646	4	30366	5	16485	6	12267	7	10900	7	92664	3
total	41362	3	53547	3	51512	3	60662	3	20588	4	227671	1

calculated by unit of reference: production region

NFI4b

biomass above ground of live trees : conifers/broadleaves · biogeographical region
unit: Mio kg
unit of evaluation: accessible forest without shrub forest
grid: grid NFI4 2009-2013
state 2009/13

	biogeographical region													
	Jura		Plateau		Northern Alps		Western Central-Alps		Eastern Central-Alps		Southern Alps		overall	
conifers/broadleaves	Mio kg	± %	Mio kg	± %	Mio kg	± %	Mio kg	± %	Mio kg	± %	Mio kg	± %	Mio kg	± %
confers	18855	5	26540	4	51116	3	12072	6	16640	5	9918	8	135140	2
broadleaves	21206	5	33963	5	23924	5	1274	16	1115	20	11095	7	92578	3
total	40061	3	60503	3	75040	3	13346	6	17756	5	21013	4	227718	1

calculated by unit of reference: biogeographical region

Practical solutions towards compatibility

- separate poststratified estimate for each set of domains (Swiss NFI):

$$\hat{T}_D^{(E)} = \sum_{e=1}^{E} \hat{T}_e \neq \sum_{f=1}^{F} \hat{T}_f = \hat{T}_D^{(F)}$$

  totals not compatible, precise estimates for domains
- intersect the sets of regularly used domains for poststratified estimation (Romanian NFI) - formula with two sets of domains:

$$\hat{T}_D^{(E)} = \sum_{e=1}^{E} \hat{T}_e = \sum_{e=1}^{E} \sum_{f=1}^{F} \lambda_{ef} \hat{Y}_{ef} = \sum_{f=1}^{F} \sum_{e=1}^{E} \lambda_{fe} \hat{Y}_{fe} = \sum_{f=1}^{F} \hat{T}_f = \hat{T}_D^{(F)}$$

  compatible totals, precise estimates for domains, certain risk to get, in general, small post-sampling strata
- estimates unconditional of known domain sizes (Czech NFI):

$$\hat{T}_D^{(E)} = \sum_{e=1}^{E} \hat{T}_e = \sum_{e=1}^{E} \lambda_C \sum_{j \in e} y_j = \sum_{f=1}^{F} \lambda_C \sum_{j \in f} y_j = \sum_{f=1}^{F} \hat{T}_f = \hat{T}_D^{(F)}$$

  compatible estimates, less precise domain estimates, but under systematic sampling the loss in precision turns out to be small

Conclusions

- eye on terminology: sampling strata, post-sampling strata, geographic domains of known size (surface area), spatial (attribute) domains of unknown size (surface area), subpopulations (e.g. of trees)
- conditional estimates for geographic domains of known surface area are precise, but conditional estimation with different sets of geographic domains leads to numerically different overall totals
- to guarantee compatibility, consider unconditional domain estimation, and to minimise losses in the estimation for geographic domains of known surface area and shape, evaluate and make use of the relatively low sample size variance under systematic sampling
- in the estimation of attribute domains of unknown size (surface area), try to identify a surrounding geographic domain of known size, such that the relative size of the attribute domain of unknown size with respect to the domain of known surface area is as large as possible