



Optimeret NO_3^- bestemmelsesstrategi

Philipp Trénel, 12-06-2015



Se 'European Agricultural Fund for Rural Development' (EAFRD)

Formål:

- At bestemme det optimale antal stik-prøver til bestemmelse af en marks NO_3^- niveau (i kg N/ha) som funktion af markens størrelse (i ha) og prøvetagningsstrategi {tilfældig sampling, grid sampling, transsekt sampling} med en tilsigtet usikkerhed på $u=\{5, 10, 15, 20, 25\}\%$ på baggrund af NO_3^- data indsamlet i et 25×25 m grid fra to danske marker.
- At bestemme laboratorieusikkerheden i bestemmelsen af NO_3^- niveauet.

Overordnede konklusioner:

Laboratorieusikkerheden i bestemmelsen af NO_3^- niveauet blev estimeret til at være på 0.64 (0.40, 1.57; 95% konfidensinterval) kgN/ha svarende til 11,5% (7,2; 28,1) af NO_3^- niveauet i de foreliggende prøver fra de to lokaliteter (Tabel 1). Der skal dog understreges, at både NO_3^- niveauet og usikkerheden i bestemmelsen af NO_3^- niveauet var signifikant forskellige på de to lokaliteter ($p < 0.05$).

Ved en tilsigtet usikkerhed vedr. sampling-strategien på 5% viste ingen af sampling-strategierne uafhængighed af markstørrelsen (Fig. 5). Grid-samplingen viste sig mindst afhængig af markstørrelsen, dog først ved usikkerheder under 15%.

Grid-samplingen resulterede generelt i færre antal stik, der skal til for at opnå en ønsket NO_3^- præcision (Tabel 3).

Det fundne mønster skyldes formentlig, at gridsampling minimerer risikoen for prøvetagning af rummeligt korrelerede prøver. Tilfældig sampling i et todimensionelt rum kan by chance sample nabo-prøver der er højt korrelerede. Transsekt sampling er intermediær, idet prøvernes afstand langs transsekten er maksimeret by design.

Det understreges her, at de her fundne konklusioner er baseret på data fra to marker, og en generalisering af konklusionerne ville kræve et større datagrundlag.

Metoder og materialer:

- *Datagrundlag:*
 - Dataset 1: NO_3^- data indsamlet i et 25×25 m grid fra to danske marker (Mark 49-0 og 41-0, foreliggende som *.xlsx filer), deres markpolygoner og kørespor (foreliggende som *.shp filer), se Fig. 1 og 2.
 - Dataset 2: Et excel dataset med i alt 232 NO_3^- , NH_4^+ og $\text{N}_{\min}$ genbestemmelser fra 2 forskellige lokaliteter analyseret på laboratoriet mellem årene 2010 og 2015. Dette data set tjener til bestemmelse af laboratorieusikkerheden.
- *Bestemmelsesusikkerheden:*

- består af to komponenter: 1) usikkerheden forbundet med samplingsstrategien til indsamling af rummeligt korrelerede observationer, $SD(\widehat{NO_3^-})$ og 2) laboratorieusikkerheden til kemisk bestemmelse af NO_3^- ud fra de poolede stikprøver, $SD(NO_3^-)_{lab}$. Den samlede bestemmelsesusikkerhed er beregnet som $\sqrt{SD(NO_3^-)_{lab}^2 + SD(\widehat{NO_3^-})^2}$ under antagelse om, at usikkerhedskomponenterne $SD(\widehat{NO_3^-})$ og $SD(NO_3^-)_{lab}$ er uafhængige af hinanden.
- *Laboratorieusikkerheden:*
 - Laboratorieusikkerheden $SD(NO_3^-)_{lab}$ blev bestemt som $SD(NO_3^-)_{lab} = \sqrt{\sum_{i=2}^I (\sum_{j=1}^J var(NO_3^-_{i,j}) / J) / I}$, hvor j og i indekserer det j 'te år og den i 'te lokalitet. Der blev på samme måde beregnet $SD(NO_3^-)_{lab}$ for hver lokalitet for sig. Der er fundet signifikant forskel i $SD(NO_3^-)_{lab}$ mellem lokaliteterne ($p=0.001438^{**}$), med højere varians i Tåstrup. De beregnede laboratorieusikkerheder fremgår af tabel 1.
- *Stokastisk simulering:*
 - Der er brugt stokastisk simulering til at estimere usikkerheden som funktion af antal stikprøver for de tre prøvetagningsstrategier "tilfældig sampling", "grid sampling", "transsekt sampling". Simuleringen blev gennemført for et sæt af marker med forskellig størrelse.
 - Forskellige markstørrelser blev generet ved at bruge de to oprindelige marker (Mark 49-0 og Mark 41-0, Fig. 1 og 2) og marker genereret ved at halvere hhv. firdele de oprindelige marker. Det samlede sæt af markerne bestod af 17 marker og fremgår af Fig. 3. Markstørrelserne varierer mellem 2.3 og 15.8 ha.
 - **Grid sampling** (Fig. 1 og 2) sigter på at simulere en jævn fordeling af stikprøverne i marken og er foregået på følgende måde:
Vælg afstand a (i m) mellem punkterne g i griddet G .
Placer G over mark M , så at g_1 i G er sammenfaldende med punkt s_1 i det fulde stikprøvesæt S for mark M .
Forskyd G i både længdegrads- og bredegradsretning med fraktionen d (i m), hvor d er trukket fra sættet $\{1, \dots, a\}$ med uniform sandsynlighed.
Find de elementer af G som er indenfor (convex hull af) området af det fulde stikprøvesæt S for mark M ; griddet kaldes nu G' og indeholder n elementer.
Tilordne til hvert element g' af G' det element s af S , som har den mindste euklidiske afstand til g' . Det resulterende sæt S'_{grid} er stikprøvesættet af NO_3^- samples med dets n elementer med en tilnærmelsesvist homogen spatial fordeling i mark M .
Beregn $\widehat{NO_3^-}_{grid}$ og $CV\widehat{NO_3^-}_{grid}$ som hhv. middel NO_3^- og $CVNO_3^-$ for subsample sættet S'_{grid} .
 - **Tilfældig sampling** (Fig. 1 og 2) er foregået på følgende måde:

Træk n elementer (uden tilbagelægning) fra sættet S med uniform sandsynlighed; det resulterende sæt S'_{ran} er stikprøvesættet af NO_3^- samples, der er tilfældigt udvalgte.

Beregn $\widehat{\text{NO}_3^-}_{ran}$ og $\widehat{\text{CVNO}_3^-}_{ran}$ som hhv. middel NO_3^- og CVNO_3^- for subsample sættet S'_{ran} .

- **Transsekt sampling** (Fig. 1 og 2) er forgået på følgende måde:

Bestem længden L_k for alle af markens M kanter k (ud fra markpolygonen).

Træk transsektens startkant k_{start} fra sættet af alle markens kanter, hvor en kants sandsynlighed for at blive trukket er givet ved $L_k / \sum L$.

Træk transsektens vinkel til kanten k_{start} tilfældigt og beregn vinkel til kørespor (givet som linjer i en *.shp fil) og transsektens længde (i m) indenfor marken M .

Gør dette indtil der er fundet en transsekt T der opfylder følgende sæt B af betingelser, dog maksimum $n_t = 1000$ gange:

- 1) Vinklen til kørespor skal ligge i intervallerne $\{\geq 20^\circ, \leq 70^\circ\}$ eller $\{\geq 110^\circ, \leq 160^\circ\}$.
- 2) Afstand til markkant ≥ 25 m.
- 3) Linjens længde min 100 m

Diskretiser linjen T til en vektor bestående af transsektpunkterne t med 1 m afstand mellem punkterne, og vælg subsættet T' af størrelse n , hvor punkterne t' i T' er ligeligt fordelt langs T . Tilordne derefter til hvert element t' af T' det element s af S , som har den mindste euklidiske afstand til t' . Det resulterende sæt S'_{trans} er stikprøvesættet af N_{min} samples med dets n elementer med en tilnærmelsesvist lige afstand langs transsekten T , der er udlagt tilfældigt i mark M under begrænsningerne B .

Beregn $\widehat{\text{NO}_3^-}_{trans}$ og $\widehat{\text{CVNO}_3^-}_{trans}$ som hhv. middel NO_3^- og CVNO_3^- for subsample sættet S'_{trans} .

- **Simuleringen** fulgte følgende algoritme:

For hver mark $M \in M = 1, \dots, 17$

beregn $\widehat{\text{NO}_3^-}_M$, $\widehat{\text{CVNO}_3^-}_M$ og markens areal A (i ha)

og for hver afstand $a \in a = 25, \dots, 500$

og for hver iteration $i \in i = 1, \dots, N=500$

Grid sample, bestem n_{grid} og beregn $\widehat{\text{NO}_3^-}_{M,n,i,grid}$ og $\widehat{\text{CVNO}_3^-}_{M,n,i,grid}$

Foretag tilfældig sampling med $n=n_{grid}$ og beregn $\widehat{\text{NO}_3^-}_{M,n,i,ran}$ og

$\widehat{\text{CVNO}_3^-}_{M,n,i,ran}$

Foretag transsekt sampling med $n=n_{grid}$ og beregn $\widehat{\text{NO}_3^-}_{M,n,i,trans}$ og

$\widehat{\text{CVNO}_3^-}_{M,n,i,trans}$

Beregn $\Delta \text{NO}_3^-_{M,n,i} = \widehat{\text{NO}_3^-}_{M,n,i} - \widehat{\text{NO}_3^-}_M$ for grid sampling, tilfældig

sampling og transsektsampling, og tilsvarende $\Delta \text{CVNO}_3^-_{M,n,i} =$

$\widehat{\text{CVNO}_3^-}_{M,n,i} - \widehat{\text{CVNO}_3^-}_M$

For hver M og n kombination beregn usikkerheden i $\overline{N03^-}$ bestemmelsen som

$$SD(\widehat{\overline{N03^-}})_{m,n} = \sqrt{\sum_{i=1}^{N=500} (\Delta N03^-_{M,n,i})^2 / N} \text{ og tilsvarende i } CVN03^- \text{ bestemmelsen}$$

$$\text{som } SD(CVN03^-)_{m,n} = \sqrt{\sum_{i=1}^{N=500} (\Delta CVN03^-_{M,n,i})^2 / N}.$$

- Output fra simuleringen for alle oprindelige og simulerede marker blev statistisk analyseret i følgende model: Bestemmelsesusikkerheden ($SD(\widehat{\overline{N03^-}})$) udtrykt i % af markens gns. $N03^-$ blev modelleret som funktion af markens areal, antal stikprøver og samplingsstrategien under hensyntagen af mark-til-mark-variationen:

$$\log \left(SD \left(\widehat{\overline{N03^-}} \right) \right) = \beta_0 + \beta_1 \sqrt{\text{antal stik}} + \beta_2 \sqrt{\text{Areal}} + \beta_3 \text{strategi} + \beta_4 \sqrt{\text{antal stik}} \times \sqrt{\text{Areal}} + \beta_5 \sqrt{\text{antal stik}} \times \text{strategi} + \beta_6 \sqrt{\text{Areal}} \times \text{strategi} + \beta_7 \sqrt{\text{antal stik}} \times \sqrt{\text{Areal}} \times \text{strategi} + e, \text{ hvor } \text{mark} \sim N(0, \sigma^2_{\text{mark}}) \text{ og } e \sim N(0, \sigma^2).$$

På samme måde blev hver mark analyseret for sig selv.

- Alle analyser er foretaget i R, R Development Core Team (2015).

Forkortelser:

$\overline{N03^-}$	middel $N03^-$
SD	standard afvigelse
$CVN03^-$	coefficient of variation, $CVN03^- = SDN03^- / \overline{N03^-} \times 100$
$\widehat{\overline{N03^-}}$	Estimeret middel $N03^-$ på basis af subsample sættet S'
$\widehat{CVN03^-}$	Estimeret $CVN03^-$ på basis af subsample sættet S'

Resultater:

Tabel 1: Laboratoriebestemmelsesusikkerheder. Se tekst for beregningsformlerne. Der er signifikant forskel mellem lokaliteternes SD(NO_3^-) ($p=0.0067^{**}$), men ikke årene ($p=0.065$). SD: standardafvigelse.

Lokalitet	År	NO_3^-	SD	NH_4^+	SD	$N_{\min}$	SD (95% c.i.)	N
Kasted	2010	2,76	0,16	15,54	0,59	18,30	0,64	12
Kasted	2011	2,31	0,27	16,94	0,67	19,23	0,71	21
Kasted	2012	2,00	0,22	16,67	1,3	18,64	1,34	24
Kasted	2013	1,97	0,34	17,37	0,59	19,34	0,84	21
Kasted	2014	2,59	0,24	17,87	0,86	20,42	0,99	20
Kasted	2015	1,48	0,33	17,76	1,29	19,21	1,29	18
<i>Kasted</i>	<i>Alle år</i>	2,15	0,48	17,10	1,16	19,23	1,03 (0,64; 2,52)	116
Tåstrup	2010	8,78	0,54	33,26	1,77	42,03	1,71	12
Tåstrup	2011	8,31	0,99	33,40	2,11	41,70	2,49	21
Tåstrup	2012	8,53	0,74	34,81	2,64	43,31	2,76	24
Tåstrup	2013	9,31	1,12	36,31	4,37	45,61	4,99	21
Tåstrup	2014	10,61	0,8	38,31	2,62	48,88	3,00	20
Tåstrup	2015	8,56	2,14	35,18	4,02	43,71	4,44	18
<i>Tåstrup</i>	<i>Alle år</i>	9,02	1,39	35,33	3,50	44,35	3,49 (2,18; 8,55)	116
<i>Begge lokaliteter</i>	<i>Alle år</i>	5,58	0,64 (0,40; 1,57)	26,21	1,65 (1,03; 4,04)	31,78	2,57 (1,82; 4,36)	232
			CV=11,5% (7,2; 28,1)		CV=6,3% (3,9; 15,4)		CV=8,1% (5,7; 13,7)	

Tabel 2: ANOVA tabel for en statistisk analyse af bestemmelsesusikkerheden ($\widehat{SD(\text{NO}_3^-)}$) udtrykt i % af markens gns. NO_3^- som funktion af markens areal, antal stikprøver og samplingsstrategien under hensyntagen af mark-til-mark-variationen: $\log\left(\widehat{SD(\text{NO}_3^-)}\right) = \beta_0 + \beta_1\sqrt{\text{antal stik}} + \beta_2\sqrt{\text{Areal}} + \beta_3\text{strategi} + \beta_4\sqrt{\text{antal stik}} \times \sqrt{\text{Areal}} + \beta_5\sqrt{\text{antal stik}} \times \text{strategi} + \beta_6\sqrt{\text{Areal}} \times \text{strategi} + \beta_7\sqrt{\text{antal stik}} \times \sqrt{\text{Areal}} \times \text{strategi} + e$, hvor $\text{mark} \sim N(0, \sigma_{\text{mark}}^2)$ og $e \sim N(0, \sigma^2)$.

	numDF	denDF	F-value	p-value
(Intercept)	1	3492	4945.959	<.0001
sqrtN	1	3492	18272.862	<.0001
sqrtAreal_ha	1	24	60.221	<.0001
strategi	2	3492	233.160	<.0001
sqrtN:sqrtAreal_ha	1	3492	600.125	<.0001
sqrtN:strategi	2	3492	11.679	<.0001
sqrtAreal_ha:strategi	2	3492	1.083	0.3386
sqrtN:sqrtAreal_ha:strategi	2	3492	0.051	0.9505

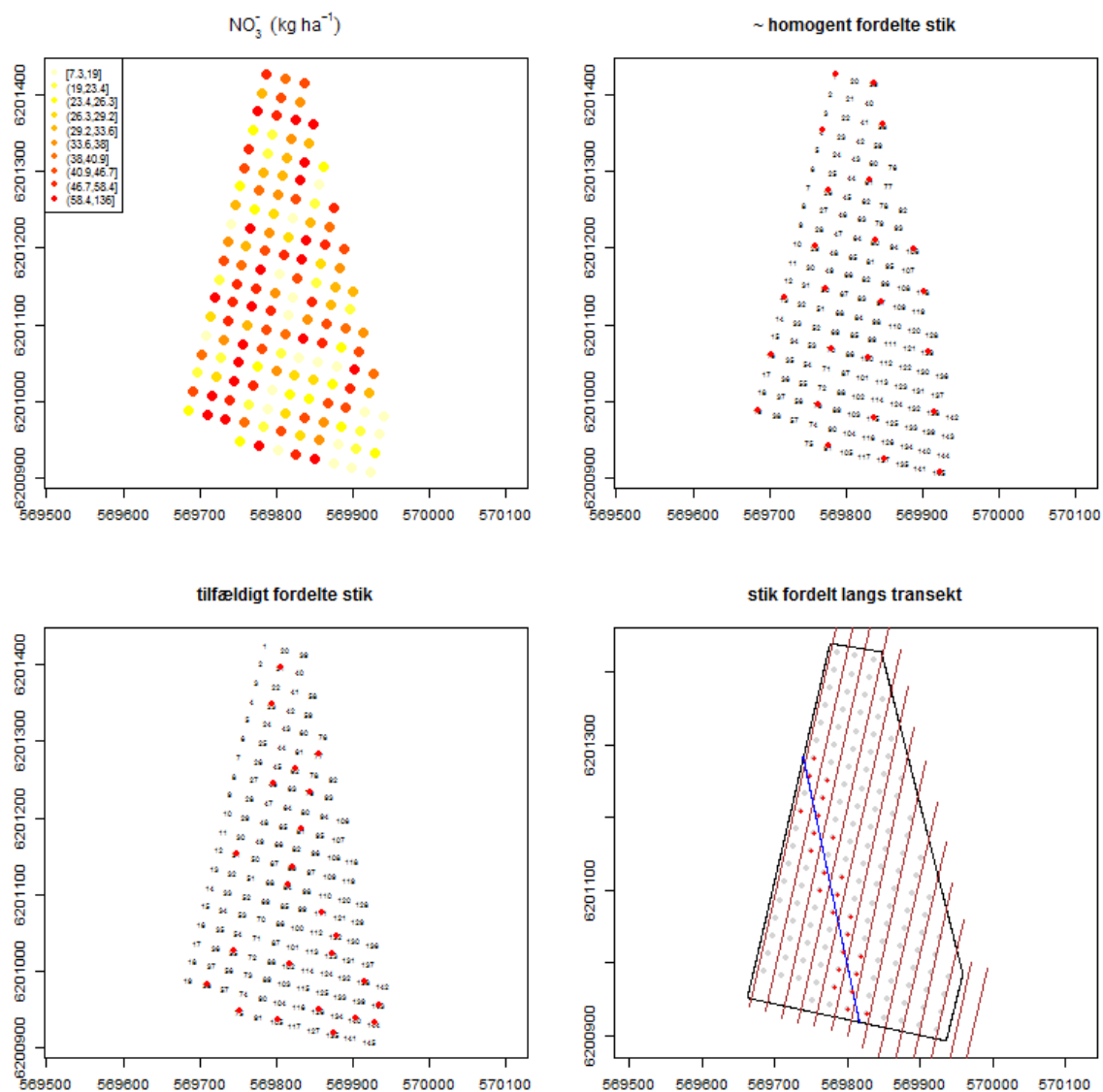


Fig. 1: Øverst til venstre: NO_3^- stikprøver fra 25×25 m griddet (det fulde sæt S) for Mark 49-0. Øverst til højre: Som før, men nu med rødt markeret de stikprøver der blev samplet i én af de 500 iterationer af grid sampling (S'_{grid}) med $n=24$. Nederst til venstre: Som før, men for tilfældig sampling (S'_{ran}). Nederst til højre: Som før, men for transekt sampling (S'_{trans}); kørespor som brune linjer; markpolygonet som sorte kanter.

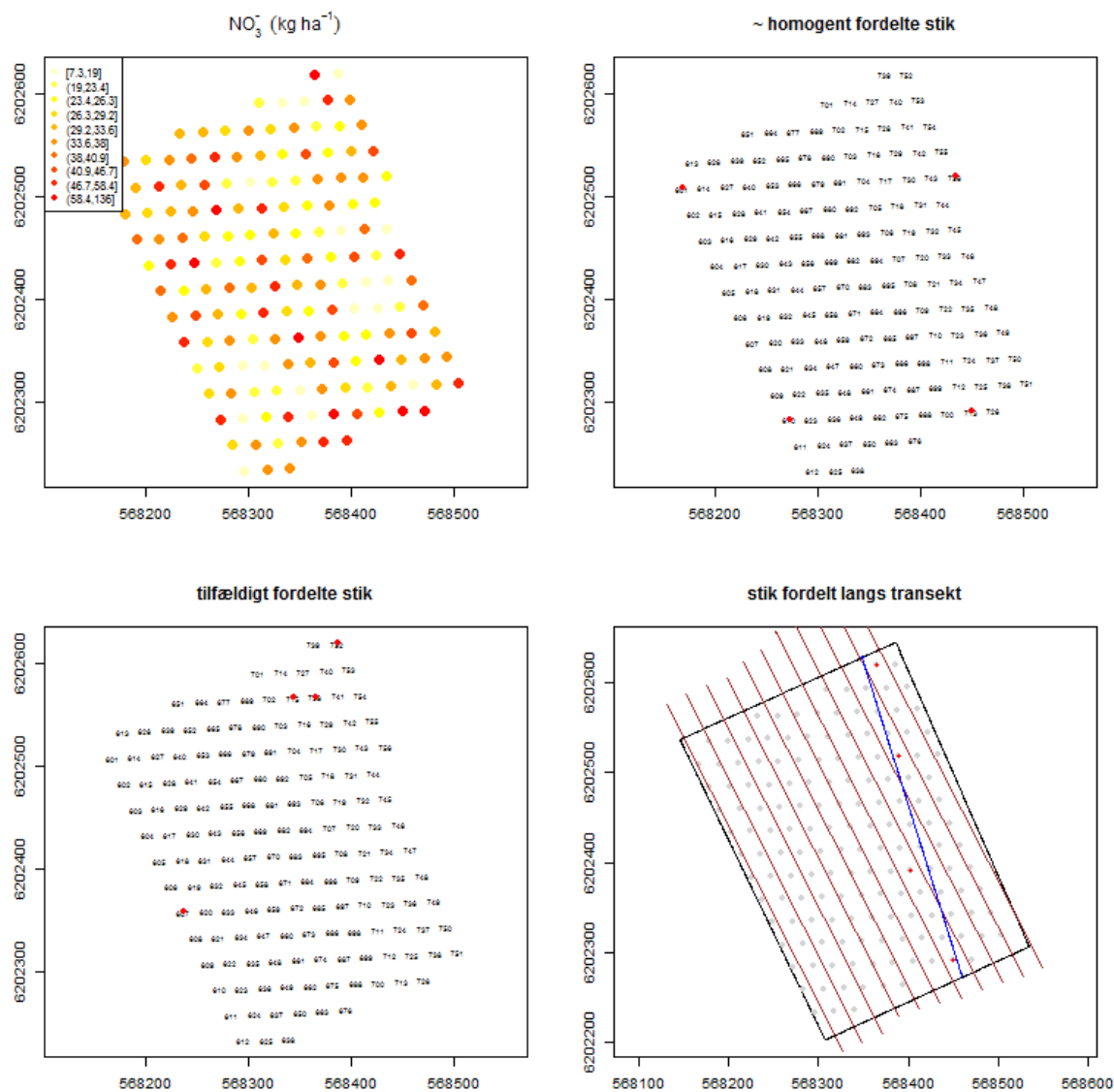


Fig. 2: Øverst til venstre: NO_3^- stikprøver fra 25×25 m griddet (det fulde sæt S) for Mark 41-0. Øverst til højre: Som før, men nu med rødt markeret de stikprøver der blev samplet i én af de 500 iterationer af grid sampling (S'_{grid}) med $n=4$. Nederst til venstre: Som før, men for tilfældig sampling (S'_{ran}). Nederst til højre: Som før, men for transekt sampling (S'_{trans}); kørespor som brune linjer; markpolygonet som sorte kanter.



Fig. 3: Sættet af markerne og tilhørende stikprøvesamples brugt i indeværende analyse. M41-0 (0) og M49-0 (0) er de to oprindelige marker; de resterende marker er simulerede marker ved hjælp af opdeling af de to oprindelige marker.

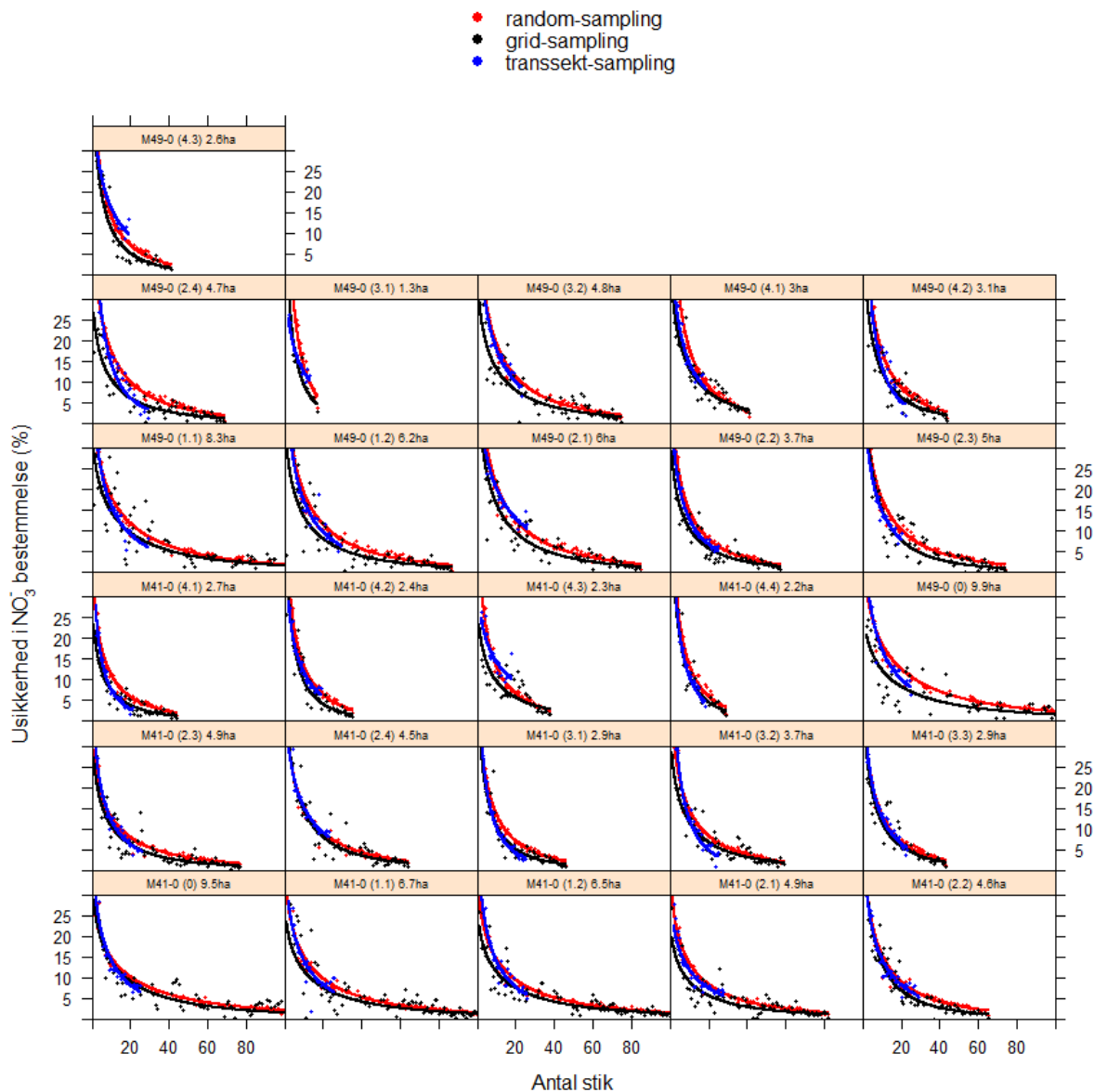


Fig. 4: NO_3^- bestemmelsesusikkerhed ($\text{SD}(\widehat{\text{NO}_3^-})$) udtrykt i % af markens gns. NO_3^- som funktion af antal stikprøver for hhv. grid sampling strategien, tilfældig sampling og transekt sampling for alle marker. Linjer viser for hvert panel den modellerede funktion $\log(\text{SD}(\log(\widehat{\text{NO}_3^-})) = \beta_0 + \beta_1 \sqrt{\text{antal stik}} + e$, hvor $e \sim N(0, \sigma^2)$.

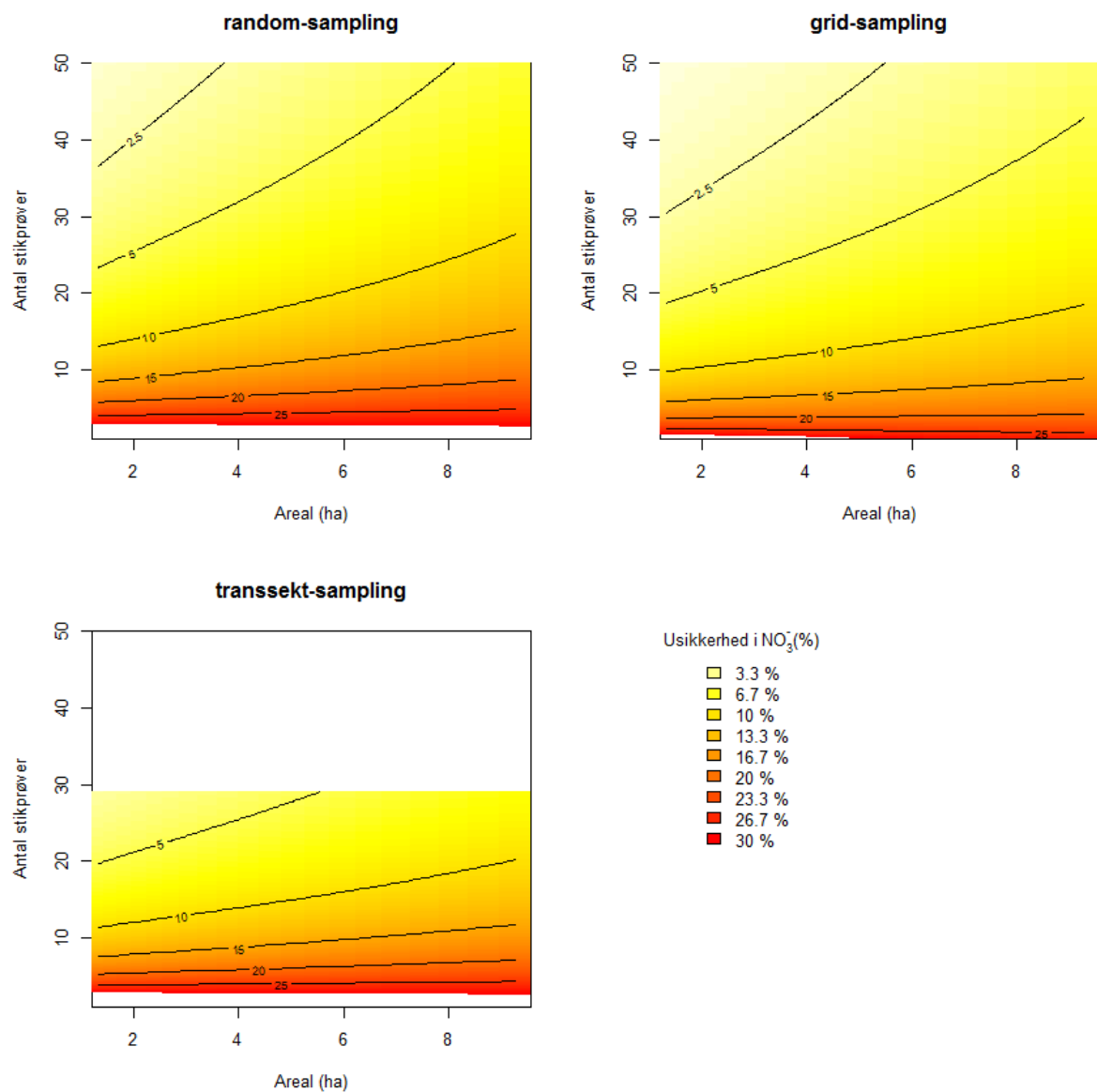


Fig. 5: Antal stikprøver som funktion af markens areal (ha) ved stigende NO_3^- bestemmelsesusikkerheder ($\text{SD}(\widehat{\text{NO}_3^-})$) udtrykt i % af markens gns. NO_3^- for hhv. grid sampling strategien, tilfældig sampling og transekt sampling. Baseret på data fra både mark 49-0 og 41-0, se tabel 2.

Tabel 3: Gns. antal stikprøver og gns. markarealet (samt deres 95% kvantiler), der har resulteret i en prædiktion af NO_3^- bestemmelsesusikkerheden prædefinerede niveauer udtrykt i % af markens gns. NO_3^-

Usikkerhed i NO_3^- bestemmelse (%)	Strategi	Antal stik	95%-nedre grænse	95%-øvre grænse	Areal (ha)	95%-nedre grænse	95%-øvre grænse
5	random-sampling	35	23	49	4,5	1,3	8,1
	grid-sampling	30	19	45	5,2	1,3	9,3
	transsekt-sampling	29	20	41	5,1	1,3	9,3
10	random-sampling	20	13	28	5,1	1,3	9,3
	grid-sampling	14	10	19	5,1	1,3	9,3
	transsekt-sampling	15	11	21	5,0	1,3	9,3
15	random-sampling	12	9	16	5,1	1,3	9,3
	grid-sampling	7	6	9	5,1	1,3	9,3
	transsekt-sampling	9	8	12	5,0	1,3	9,3
20	random-sampling	7	6	9	5,0	1,3	9,3
	grid-sampling	4	4	4	4,9	1,3	9,3
	transsekt-sampling	6	5	7	4,9	1,3	9,3
25	random-sampling	4	4	5	4,9	1,3	9,3
	grid-sampling	2	2	2	4,7	1,3	9,3
	transsekt-sampling	4	4	4	4,8	1,5	9,3

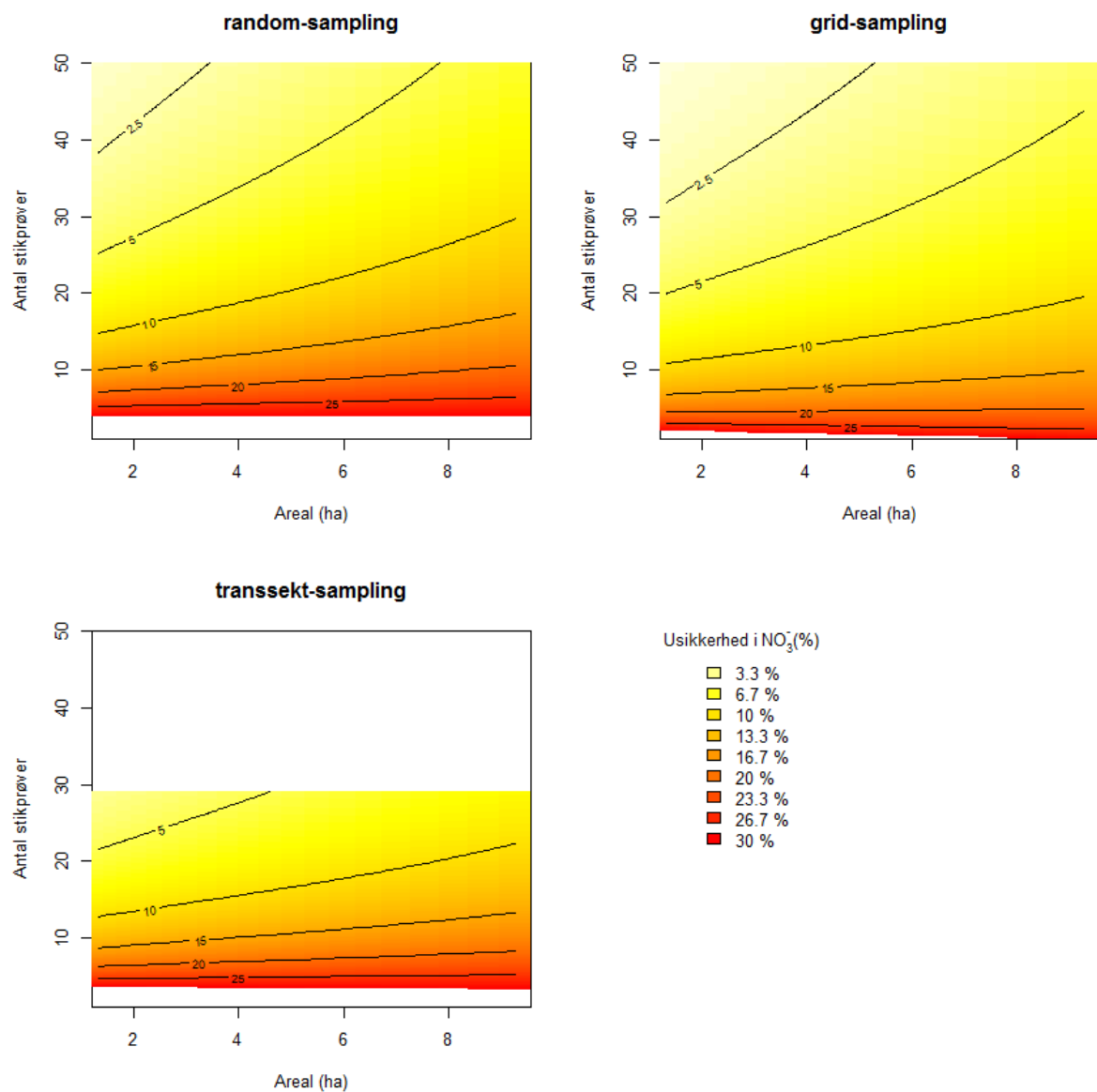


Fig. 6: Antal stikprøver som funktion af markens areal (ha) ved stigende NO_3^- bestemmelsesusikkerheder ($\text{SD}(\widehat{\text{NO}_3^-})$) udtrykt i % af markens gns. NO_3^- for hhv. grid sampling strategien, tilfældig sampling og transekt sampling. Baseret på data fra kun mark 49-0.

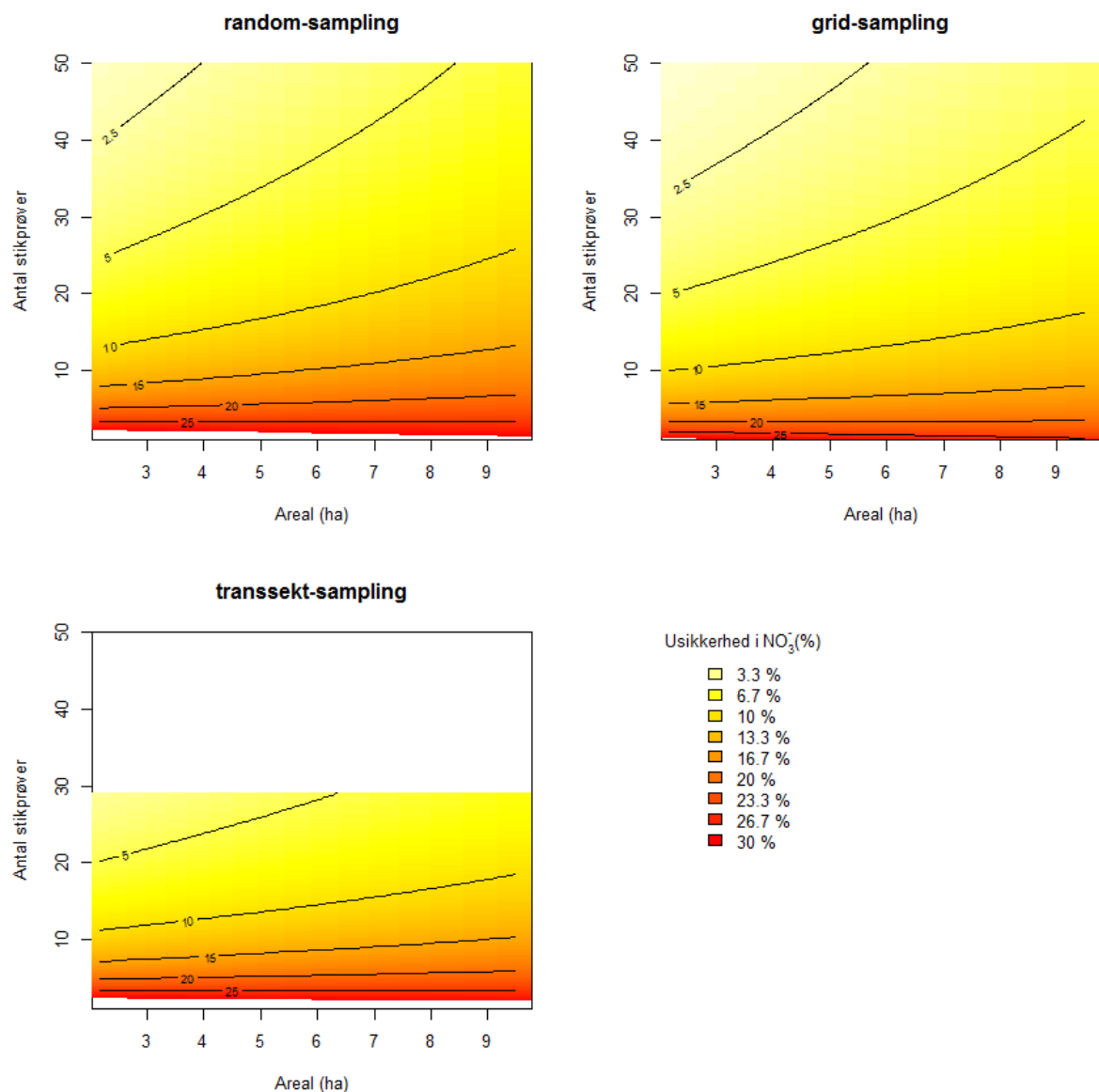


Fig. 7: Antal stikprøver som funktion af markens areal (ha) ved stigende NO_3^- bestemmelsesusikkerheder ($\text{SD}(\widehat{\text{NO}_3^-})$) udtrykt i % af markens gns. NO_3^- for hhv. grid sampling strategien, tilfældig sampling og transekt sampling. Baseret på data fra kun mark 49-0.

Referencer:

R Development Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.