Dataanalys kopplat till undersökningar

Seminarium om undersökningsmetoder för förorenade områden, Malmö 6-7 maj

Jenny Norrman,
SGI, Chalmers FRIST

Innehåll

- Inledning
- Provtagningssyften
- Terminologi
- Konceptuell modell
- Provtagningsstrategier
- Provtagningsosäkerhet
- Hur många prover?
- Datautvärdering
- Programvaror

Inledning

- Varför provtar man?
  - Del i beslutsunderlag
  - Forskning
- Viktiga frågor att ställa sig:
  - Vad är provtagningssyftet?
  - Hur skall resultaten från provtagningen användas?
- Enligt Bättre Markundersökningar (SGI, 2006 – 14 rapporter) är det ofta brister i hur man preciserar provtagningssyftet.

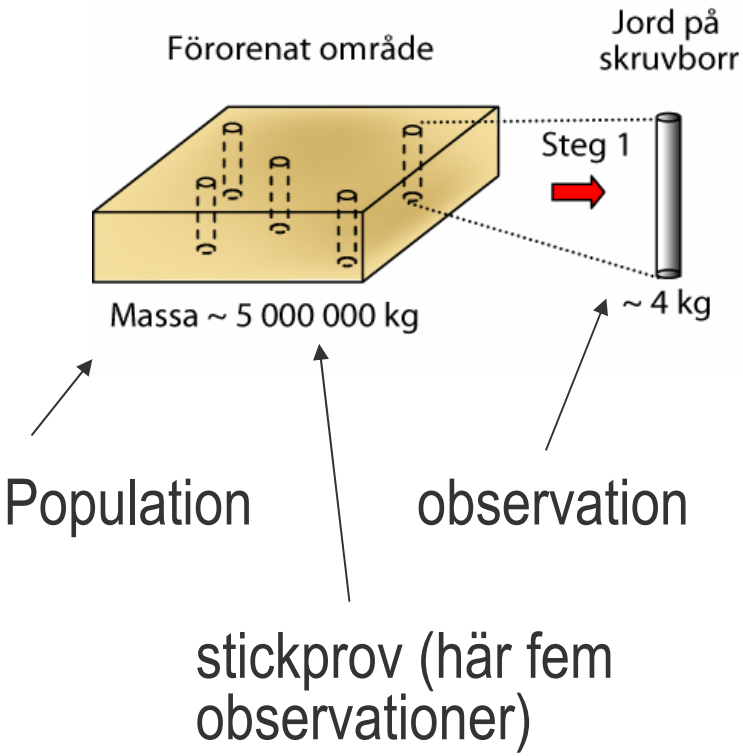
Provtagningssyften

- Vad finns det?
- Hur mycket finns det?
- Var finns det?
- Sker det någon förändring över tiden?
- Kombinationer av ovanstående…

Terminologi

- Population och variabel
- Stickprov:
  - Statistisk definition: n stycken observationer
  - Ofta i fält: ett insamlat prov
- Support och representativitet
- Samlingsprover
- Rumslig korrelation

Population, variabel, observation, stickprov

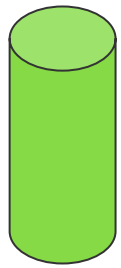


Variabel, t ex halten i en observation

- Population: en grupp individer eller ett antal fysiska objekt
- Individerna eller de fysiska objekten har en eller flera egenskaper av intresse, vilka är observerbara storheter eller *variabler*

Support och representativitet

- ”Sample support” – strikt definierat som provets volym, form och orientering



- Representativitet – ett uttryck för vilken grad ett prov kan användas för att karaktärisera den population som ska undersökas i förhållande till det beslut som skall fattas

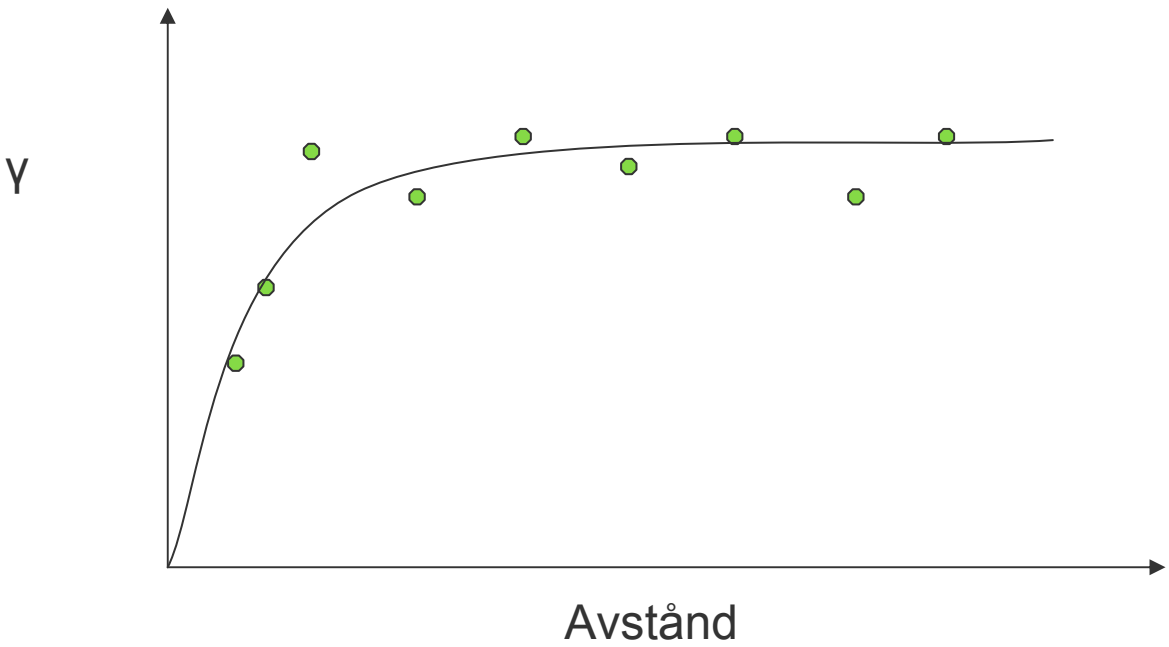
Samlingsprov

Man använder sig av delprov i en större volym för att skatta medelkoncentrationen

- Fördelar:
  - Färre analyser, dvs. mindre kostsamt
  - Större support, dvs. mindre variabilitet i data
- Nackdelar:
  - Mindre information om variabilitet
  - Stora provvolymer som kan påverka labkostnaden

Rumslig korrelation

- Graden av beroende mellan två punkter i rummet
- Modelleras ofta mha variogramanalys



Konceptuell modell

- Vad är en konceptuell modell?
  - Sammanfattning av all tillgänglig kunskap, både kvalitativt och kvantitativt, av områdets föroreningssituation

- Hur används den?
  - Utforma provtagningen

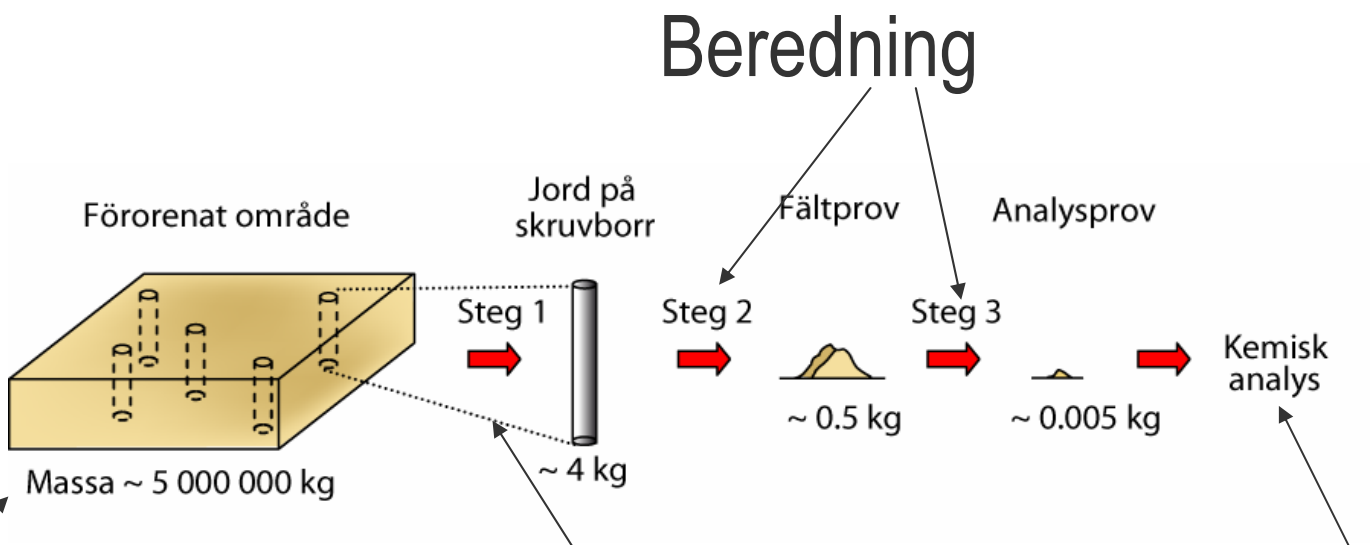
Provtagningsstrategier

- Sannolikhetsbaserad provtagning
  - Princip: varje element i en population har samma sannolikhet att bli utvald
  - Statistiska metoder kan användas för utvärdering
- Riktad provtagning
  - Princip: baserad på förhandskunskap
  - Analys med statistiska metoder ger inte representativa resultat
- …i de flesta fall en blandning av dessa.

Slumpmässiga urval

- I små populationer är det möjligt att göra en totalundersökning.
- I stora populationer är det sällan möjligt.
- Då får man göra ett urval och dra slutsatser om hela populationen på basis av det begränsade urvalet.
- Eftersom man drar slutsatser ifrån ett begränsat urval är det bra att kunna beräkna felet i extrapolationen
- Det kan man göra om urvalet sker slumpmässigt.

Provtagningsosäkerhet



Strategier att hantera osäkerheter när man planerar provtagningen

1. Osäkerheter minimeras för en given budget
2. Kostnaden minimeras utifrån en given accepterad osäkerhet
3. Osäkerheter hanteras genom att följa ett regelverk för hur provtagning skall utföras
4. Osäkerheter ställs i relation till en potentiell risk och provtagningen planeras utifrån en balans mellan kostnaden för prover (att minska osäkerheterna) och hur stor riskkostnad osäkerheterna är förknippade med (datavärdesanalys) (Back, 2006)

Hur många prover?

- US EPA strategi: kostnaden minimeras utifrån en given accepterad osäkerhet
- Den accepterade osäkerheten kan dock variera beroende på platsspecifika förutsättningar.

		Verkligt förhållande	
		Förorenat	Ej förorenat
Tolkning av data	Ej förorenat	Typ I-fel, alfa	Korrekt beslut, 1-beta
	Förorenat	Korrekt beslut, 1-alfa	Typ II-fel, beta

Ex: Givet ett acceptabelt typ 1-fel

Antalet prov baserat på ett konfidensintervall för medelvärdet:

$$n = \left(\frac{t_{1-\alpha;n-1} \cdot s}{e} \right)^2$$

- $t_{1-\alpha;n-1}$ = från Student's t-fördelning, α acceptabelt typ I fel, n-1 frihetsgrader
- e = max acceptabelt fel i medelskattningen (bredden på konfidensintervallet)
- s = standardavvikelsen

Ex forts.

- α = 0.05
- e = 50 mg/kg
- s = ?
- Skatta genom $s \approx \frac{\max - \min}{4} = \frac{600 - 200}{4} = 100 mg/kg$
- (Back, 2004)
- Första gissning på n, t ex 11 (z-förd)

- Iterera fram:
- n = 18 prover

Ex: Givet acceptabelt typ I och II fel

Antal prover det krävs för att upptäcka en given minsta detekterbar skillnad från riktvärdet (efter Grandin, 2006):

$$n = \left(\frac{s\left(t_{1-\alpha;n-1} + t_{1-\beta;n-1}\right)}{MDS} \right)^2$$

- s = standardavvikelsen för populationen
- $t_{1-\alpha;n-1}$ från Student's t-fördelning, α acceptabelt typ I fel, n-1 frihetsgrader
- $t_{1-\beta;n-1}$ från Student's t-fördelning, β acceptabelt typ II fel
- MDS = Minsta Detekterbara Skillnad i absoluta tal.

Ex forts.

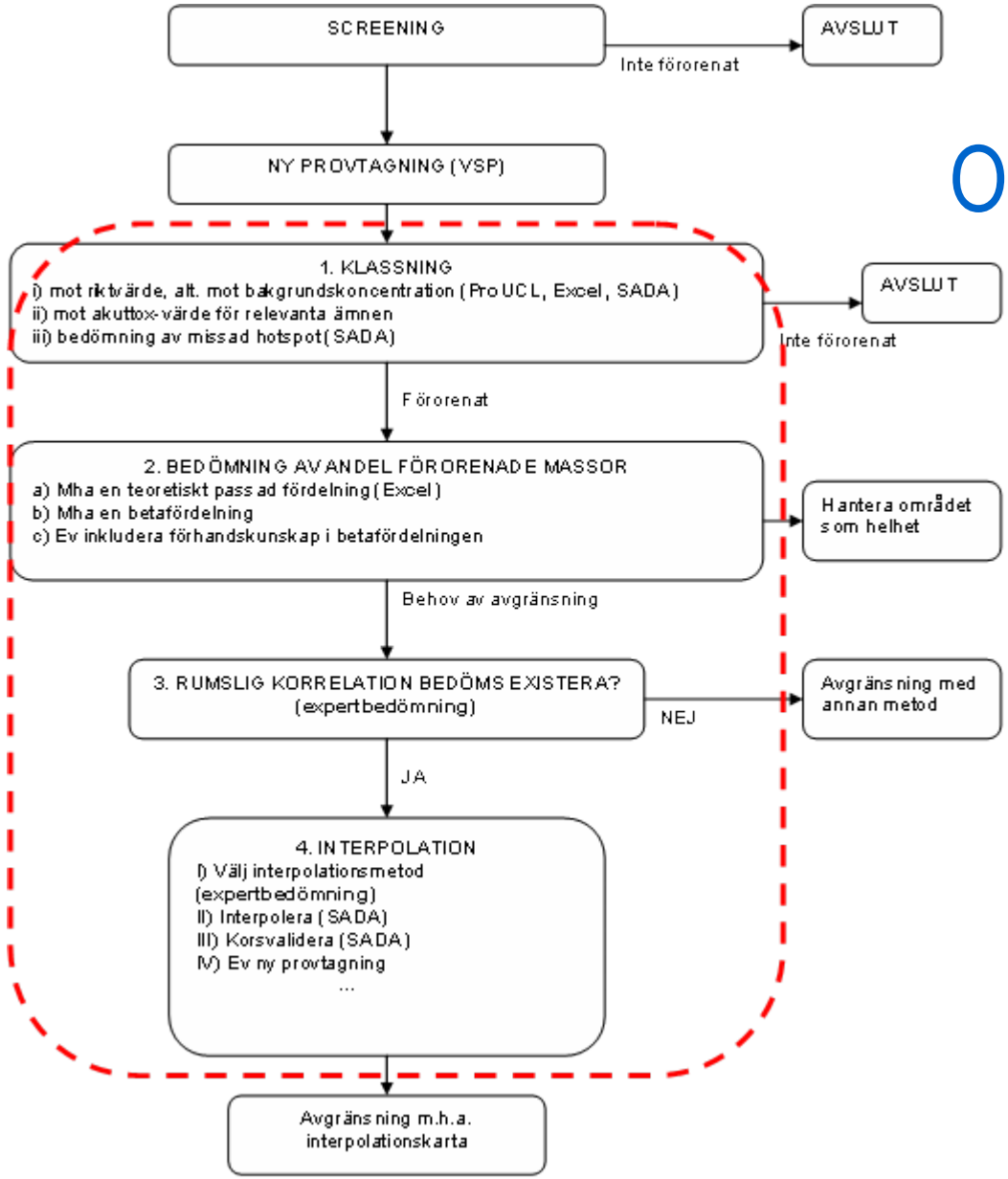
- $\alpha = 0.05$
- $\beta = 0.20$
- MDS = 50 mg/kg
- s = 100 mg/kg
- Första gissning på n, t ex 26 (z-förd)

- Iterera fram:
- n = 44 prover
- OBS! Antagen normalfördelad data!

Varför kvantifiera osäkerheter och använda statistik?

Kvantifiering:

- minskar behovet av bedömningar/tyckande.
- möjliggör någon form av vetenskaplig kritik.
- ökar (förhoppningsvis) transparensen i våra ställningstaganden.
- ökar möjligheten till oberoende jämförelser.
- ökar vår egen förståelse för olika risker.



Datautvärdering
Optimerad utvärdering

1. Klassning
2. Bedömning av andel förorenade massor
3. Rumslig korrelation
4. Interpolation

1. Klassning

- Jmf representativ parameter med riktvärden (eller bakgrundshalt)
- Jmf representativ parameter med akuttox-värden
- Bedömning av missad hotspot

- Slutsats: förorenat/ej förorenat

2. Bedömning av andel förorenade massor

- Använda en teoretisk fördelning
- Använda datas fördelning
- Använda en betafördelning
- Ev inkludera subjektiv kunskap i betafördelningen

Slutsats: behov av avgränsning eller inte

3 & 4 Rumslig korrelation och interpolation

- Bedömning av om det föreligger rumslig korrelation på en skala som är praktisk för interpolation
- T ex variogramanalys
- Slutsats: avgränsning med interpolation eller annan metod
- Deterministiska: inverse distance, triangulering…
- Geostatistiska: kriging, geostatistisk simulering
- Korsvalidering

Data verifiering, validering

- Se över provtagningssyftet
- Support?
- Riktad provtagning eller slumpmässig?
- Stratifiering?
- Data under detektionsgränsen?
- Outliers?

Sammanfattning

- Syftet med en undersökning är av avgörande betydelse för hur undersökningsstrategin bör utformas.
- En konceptuell modell (hypotes) av föroreningssituationen är avgörande för kvalitén på provtagningen och dataanalysen.
- Hur säker behöver du vara för att ta beslut?

Gratis programvaror (US EPA)

- VSP (Visual Sample Plan) – planering av undersökningar och utvärdering av data
- SADA (Spatial Analysis and Decision Assistance) – utvärdering av data, även interpolation, samt planering av undersökningar
- ProUCL – endast utvärdering av data, ej interpolation