Sparse sampling:
Spatial design for aquatic monitoring

Melissa Dobbie

CSIRO Mathematical and Information Sciences, Brisbane
AusCan Scholar visit to Canada, January-February 2008

Acknowledgements

➢Australian and Canadian Statistical Societies (SSAI and SSC) for their generous funding to provide me with an unique opportunity to meet Canadian statisticians, learn from them, tell them something about my statistical endeavours, and ultimately to enrich my statistical career.

➢CMIS for their support in enabling me to take up this scholarship.

➢My collaborators Brent Henderson (CMIS) and Don Stevens (Oregon State University)

➢**contributions to paper we have written and submitted to *Statistics Surveys* in December (which provides detail of what I'm talking about today).**

➢ **for use of some of their slides.**

Oz snapshot ….

Brisbane (*the River city*)
- Population = 2 million (i.e. a bit smaller than Vancouver)
- 3rd largest city
- Subtropical climate: mean max/min temp 25.5℃/15.7 ℃
- Mean annual rainfall = 1146 mm
- Famous Brisbanites include Steve Irwin, Greg Norman, Keith Urban, Geoffrey Rush, Kevin Rudd, Steven Bradbury.
- Some Brisbane statisticians you may have heard of: Geoff McLachlan, Tony Pettitt, Kerrie Mengersen, Bill Venables, Annette Dobson, Kaye Basford.

Queensland *(aka Sunshine State, Smart State)*
- Population = 4.2 million (i.e. a bit smaller than BC)
- Total area = 1.9 million km^2 (twice size of BC)
- 5 World Heritage listed preservation areas including Fraser Island and the GBR
- SSAI Qld Branch has around 75 full members employed by CSIRO (15%), Universities (40%), local govt depts (30%), other (15%).

Australia
- Population = 20 million (*cf* 33 million in Canada)
- SSAI has 750 members

Motivating example

Why monitor?

Aquatic monitoring occurs for many reasons

➢to measure current state or condition of the environment, including provision of important baseline information

➢to assess the impact of a particular event or intervention

➢to assess if there has been any change in important environmental variables.

➢to satisfy regulatory or compliance requirements.

➢to identify 'hotspots' across some spatial domain

Prolonged drought in Australia means knowledge about water resources (quantity, quality, where) has become top priority.

Aquatic Monitoring

Key components

Objectives — Design — Analysis — Reporting

Aquatic Monitoring

Key components

Objectives → Design ← Analysis ← Reporting

- **Our focus here is on the monitoring design**
- **Monitoring objectives drive the design**
- **Analysis & reporting needs dictate the information that the design must provide**
  - e.g. if we need to report at monthly intervals the design must deliver data that support that.
  - e.g. if we need to make spatial predictions that should be enabled by the design wherever possible

Key aspects of aquatic monitoring design

- **resource to be monitored**
- **indicators to be measured**
- **spatial location of monitoring sites** — Spatial Design
- **temporal frequency of monitoring**
- **method for obtaining the response (operational)**
- **analysis & summary methods**

Establishing the context

- Focus on freshwater stream networks
- Target populations such as streams or rivers that are present only on a linear network within a bounded area → linear resource
- Essentially 1D continua embedded in 2D space
- Comprise an uncountably infinite collection of points.
- No well-defined natural units for these types of systems
- Attributes are defined at all points of the network.

Objectives of review

- Present a synthesis of spatial design research for monitoring stream networks, and
- Document the current status of this area of research

to ultimately

- Inform development of appropriate spatial designs for future aquatic monitoring problems.

From "The Southeast Queensland drought to 2007" report

Sparse sampling???

Helps to capture the essence of important design issues that arise and need to be considered in more detail.

In particular, the description of sampling for these applications as sparse is motivated by the

- tension between the cost of sampling and the number of sites to be sampled
- allocation of resources across space and time
- unique structure of the target population within the landscape
- likelihood of increased sparse responses due to target populations less likely to be static over a sampling period
- prevalence of ephemeral streams (both naturally occuring and as a result of the prolonged drought).

Key issue:
Reliable inference demands that relationship between the sample and the population be characterized by either: model / theory or through the sampling process

Aquatic Resource
POPULATION

↓ site selection → **SAMPLE** ↑ inference or extrapolation ←

Monitoring objectives inform site selection

- Are we interested in assessing trend or status?
- Are we interested in regional assessment of a population characteristic?
- Is it important to be able to aggregate and report at a hierarchy of different scales?
- Are we interested in spatial prediction? Or estimating regression or spatial parameters?

The most appropriate site selection approach depends intimately on the monitoring objectives.

Sampling aquatic resource populations

- Often expensive and time-consuming to sample
- Typically only possible to sample a small proportion of population.
- Spatial patterns (e.g. gradients, periodic effects) often are critical
- Often interested in many responses (e.g. water quality variates)
- Involve 0, 1, and 2 dimensional populations
  - points: e.g. bore holes, farm dams
  - lines: e.g. rivers & streams
  - areas: e.g. lakes, estuaries, wetlands
- May need to sample a continuous population, e.g. a stream
- It is often difficult to define a reliable sampling frame
- Non response (e.g. inaccessible points) can be substantial

Traditional site selection methodologies

Methods for sampling environmental resources have often been fairly ad hoc and have tended to appeal to expert knowledge.

Convenience Representative

- Unknown relationship between the sample data and population characteristics
- Basis for extrapolation and inference is not clear
- Sites that are representative for one variable may not be representative for any other variable
- If sites are truly representative of average response then the extremes are suppressed.
- Most likely unable to make statistically-valid regional assessments of condition or report at a hierarchy of scales
- Increasing requirement for science to be defensible and for scientists to be environmentally accountable

Statistically-based approaches

There are three popular statistically-based philosophies for choosing the spatial monitoring design:

- Geometric
- Model-based
- Probability-based

Geometric methods

- Consider how well a set of design points covers the domain
- Design criterion is based on geometry and the distance between both current and potential sample locations
- Typically based on heuristic arguments
- Often used for exploratory purposes
- Include regular lattices, triangular networks, space-filling designs

Model-based design

- A statistical model may be used to describe the underlying environmental process → appeal to model for population to choose sites
- Relationship between the sample and the population described by model
- Precise inference may be possible from limited number of samples; Inference is based on the model
- Stochastic element is embedded in the model process
- **Issues:**
  - Reliability of the inference depends on the adequacy of the model
  - Forming the model requires either knowledge of the underlying process of interest and/or sufficient data on that process – which may be unavailable or unknown.
  - Some ecological systems may demand complex models

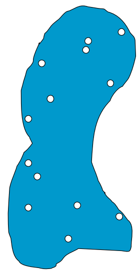
Probability sampling

- **Probability sample has 3 distinguishing features:**
  1. Population being sampled is explicitly described
  2. Every element in the population has the opportunity to be sampled with known probability
  3. Selection is carried out by a process that includes an explicit random element.
- **Randomisation particularly important as avoids bias and ensures sample is representative**
- **Also generality and validity extend from randomness**
- **Probability sampling allows estimates to be aggregated from the local to the national level**
- **Objective assessment of uncertainty available**
- **Can incorporate expert knowledge in both the design and analysis phases**
- **Primary focus is to enable us to make inferences about a relevant attribute (i.e. mean, total, quantile,etc.) for a population on some spatial domain**
  - e.g. proportion of lakes that might be considered eutrophic

Basic Spatial Survey Designs

Simple Random Sample

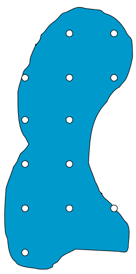
- Representative, avoids bias
- Flexible – easy to add extra points
- No assurance of spatial balance or regularity
- Variance can be high
- Inefficient if spatial structure / correlation



Basic Spatial Survey Designs

Systematic sampling

- Regular. Good spatial balance
- More precise than SRS if spatial correlation exists
- Potential inefficiencies for populations with patchy or periodic responses
- Limited flexibility to alter point density or add extra points
- Boundary problems may exist

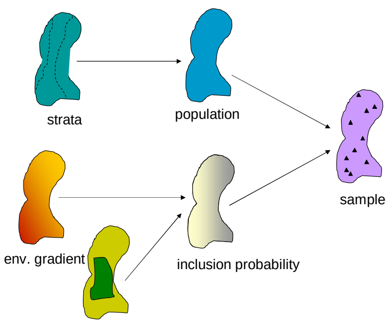


Basic Spatial Survey Designs

Stratified sampling

- Break population into strata, providing some spatial structure / balance and reducing overall variability
- Strata – polygons, natural boundaries, management zones, …
- Use SRS or systematic or other within carefully chosen strata
- Allows adjustment of sampling effort according to different sub-populations, operational / administrative efficiency & prior knowledge of the ecological population or study area
- Lower flexibility (strata specialised for a specific application lose utility for other applications)
- Not necessarily well-balanced spatially
- Small sample sizes per strata cause the greatest loss of efficiency in the presence of non-response

Sampling process



Desirable attributes of spatial survey design

- **Good spatial regularity / balance**
- **Allow for adjustment of sample sizes dynamically**
  - e.g. for non response or augment as needed
- **Accommodate variable inclusion probabilities**
  - Generalization of stratified sampling that allows selection probabilities to vary continuously, instead of being constant within discrete strata and only varying among strata.
  - e.g. account for travel time / costs
  - e.g. levels of an environmental stressor
  - e.g. target particular habitats

→ **spatially-balanced sampling**

Spatially-balanced sampling

- Combination of simple random and systematic sampling
- Guarantees all possible samples are distributed across the aquatic resource
- Flexibility to adjust point density
- Attempts to maximize spatial independence among sample locations
- Easy to implement for point, linear and areal aquatic resources

Methods

- Hierarchical quadrant recursive ordering
- Random-tessellation Stratified (RTS) design
- Generalized random-tessellation stratified (GRTS) design
- Reverse randomized quadrant recursive raster (RRQRR)

Stevens & Olsen (2004) - GRTS

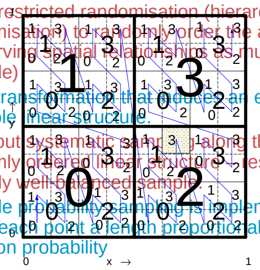
- **Developed a unified strategy for selecting spatially-balanced probability samples of natural resources**

Generalised Random-Tessellation Stratification

- **Accommodates regular issues that occur in natural resource populations**
  - variable probability
  - inaccessibility
  - missing data
  - poor frame information
  - uneven spatial pattern
  - panel structures

GRTS in a nutshell

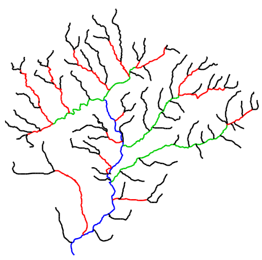
- Create a function that maps 2D space into 1D space, thereby defining ordered spatial address
- Use a restricted randomisation (hierarchical randomisation) to randomly order the address (preserving spatial relationships as much as possible)
- Apply transformation that induces an equi-probable linear structure.
- Carry out systematic sampling along the randomly ordered linear structure resulting in a spatially well-balanced sample
- Variable probability sampling is implemented by giving each point a length proportional to its inclusion probability



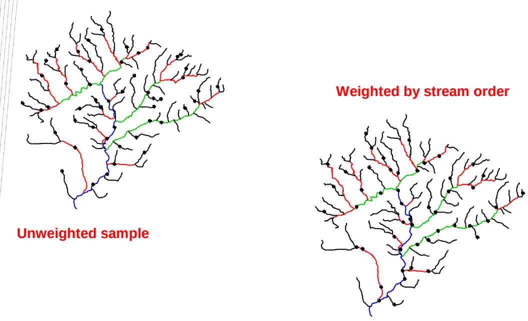
An example: sampling a stream network (thanks to Don Stevens)

- A GIS represents a stream network as a series of straight-line segments
- Assume that the resolution is high enough so that the inclusion probability density is constant on each stream segment

Sampling a Stream Network



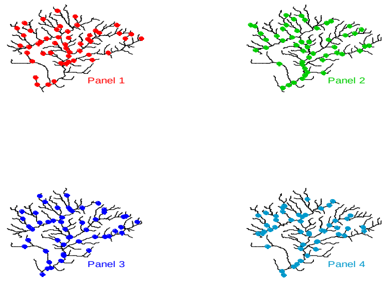
50 point sample



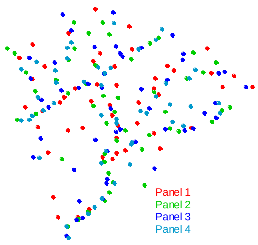
Stream order

	1	2	3	4
Percent Length	60	21	12	7
Unweighted sample	64	18	8	10
Sample weighted by stream order	40	26	18	16

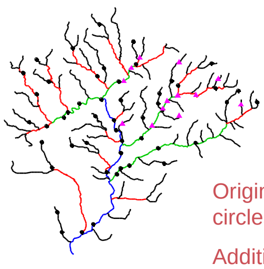
Spatio-temporal design



All panels overlayed



Augmented sample in priority area



Original points are black circles

Additional points are pink triangles

Key messages

- **Probability-based sampling underpins spatial survey sampling of aquatic resources and has been implemented in environmental surveys overseas, e.g. EMAP in U.S.**
- **Main reasons for considering them are: unbiasedness, validity of inference, clarity of analysis, aids hierarchical reporting, can include expert knowledge via inclusion probabilities…**
- **Improvements to the methodologies are a current research thread**
  - Generalized random-tessellation stratified (GRTS) design
- **Site selection challenges are appropriate to field & sensor-based sampling**

In summary ….

GRTS design approach …..

- Ensures spatial balance prevails
- Enables dynamic adjustment of sample size
  - non-response
  - imperfect sample frame formation
  - Sub-populations of interest may change over time
- Accommodates variable inclusion probability
  - Legacy sites
  - Political, economic, scientific reasons affecting site selection
- Is implemented by US EPA in their EMAP and in other countries outside of the USA
- Can be applied to monitoring 0-, 1-, or 2-dimensional (natural) resources
- Works for large-scale environmental monitoring problems
- Addresses regional, continental and global environmental issues

CSIRO Mathematical and Information Sciences
Melissa Dobbie

Email: Melissa.Dobbie@csiro.au
Web: www.csiro.au/people/Melissa.Dobbie

Thank you

Contact Us
Phone: 1300 363 400 **or** +61 3 9545 2176
Email: Enquiries@csiro.au **Web:** www.csiro.au