Why don’t we ask?

Can citizen science improve ecology and conservation?

Kerrie Mengersen
D. Professor of Statistics, QUT
Australia

With colleagues from the Bayesian Research and Applications Group (BRAG), ACEMS, Australian Institute for Marine Sciences and Queensland Government Department of Environmental Sciences
Citizen Science (CS): What?

- **1995**: Irwin (UK) opening up science and science policy processes to the public
- **2009**: Bonney et al. (US) public-participation engagement & science communication projects
- **2018**: data, expert information, verification, analysis

CS: Who?

• General public
• Educated public
• Trained public
• Stakeholders
• Experts

CS: Where?

- Biodiversity
- Birds
- Meteor science
- Earthquake science
- Taste research
- Bioindicators on a national scale
- Humans as long-distance dispersers of plants
- Land-use preferences of flower visitors


http://www.photovolcanica.com/VolcanoInfo/Yasur/Yasur.html
CS: How?

- Zooniverse.org
  over 12M+ public observations
- eBird
  hundreds of thousands of contributions
- Evolution MegaLab
  thousands of volunteers in 15 countries in Europe
- iWitness pollution map
  crowdsourcing petrochemical accident research
- Open-Phylo
  crowd-computing platform for multiple sequence alignment
- OPAL
  open-air laboratories engagement project in England


http://www.birds.cornell.edu/page.aspx?pid=1664
CS: why (not)?

Pros:
- Cost effective
- Participatory
- Timely
- Space-time range
- Broadly applicable

Cons:
- Potential bias
- Variable quality
- Perception
- Ethics
- Ownership
- Commitment
**CS: important factors**

- Vianna *et al.* (2014): a strong correlation between number of grey reef sharks observed by dive guides and telemetry data at both daily and monthly intervals; behaviour of sharks was not affected by diver presence.

- Matteson *et al.* (2012): volunteer monitoring protocol was important to estimate the number of butterfly species present in Chicago and New York.

- Jordan *et al.* (2012): volunteer participation can enhance the data generated by scientists alone; data quality is improved by training and over time.

- Sauermann and Franzoni (2015) sustainability depends on degree of interest and hence provision of continued labour inputs

CS: important factors

• Kremen et al. (2011): utility may be restricted to detection of community-level changes in abundance, richness, or similarity over space and time, abundance or frequency of occurrence of specific pollinator species or groups.

• Paul et al. (2014): wildlife-conflict reports, significant spatial agreement and robustness; should record search and reporting effect.

• Bonney et al. (2014) consistent methodology and training; with appropriate protocols, training, and oversight, volunteers can collect data of quality equal to those collected by experts.

CS: how close?

CS and Bayesian modelling

Posterior = Likelihood / c * Prior

Meld → Augment → Elicit
Design → Adjust → History matching
1. CS + Melding posterior distributions

Suppose we have a mechanistic (deterministic) model and we want to incorporate uncertainty into inputs and outputs.

• Assign a prior to both input and output.
• Calculate likelihoods for both input and output.
• “Meld” these four sources of information.

Poole and Raftery (2000)
Case study:
Revised universal soil loss equation

\[ A = R \times K \times L \times S \times C \times P \]

A: soil loss per unit area
R: rainfall and runoff factor
K: soil erodibility factor
L: slope length factor
S: slope steepness factor
C: cover and management factor
P: supporting practice factor

CS and Bayesian modelling

Posterior = Likelihood / c * Prior

Meld
Design
Augment
Adjust
Elicit
History matching
Suppose that we have a lot of citizen-contributed data.

If we have a specific question, we don’t need to analyse all of it.

Use experimental design principles to select the data required to answer the question.
A decision analysis approach to experimental design

Select optimal settings $d$ in some design space of options, $d \in D$.  

*ie* maximise expected return as quantified through a known utility function $U(d, \theta, y)$ where $y$ is a future dataset that may be observed when design $d$ is applied.

Example:

• Regression with continuous response $Y$ and measurement covariates $X$, and study objective to learn about parameters $\theta$ of a mean regression function $E[Y] = f(X; \theta)$

• Design points might be points in $X$ with $d \in D \subseteq d$

• Utility might be based on variance of unbiased estimator that targets the true unknown $\theta$. 

Bayesian experimental design

- Focus on expected gain in Shannon information from the prior to the posterior distribution (aka mutual information, Kullback-Leibler distance) or spread of posterior distribution (precision, entropy)

- Optimal design $d^*$ maximises expected utility function $U(d)$ over the design space $D$ with respect to model parameters $\theta$ and future data $y$:

$$d^* = \arg \max_{d \in D} E_\Theta U(d, \theta, y) = \int_Y \int_\Theta U(d, \theta, y) p(\theta, y|d) d\theta dy$$

$$p(\theta, y|d) \propto p(y|\theta, d)p(\theta)$$
Experimental design in the context of big data

1. **Answer questions of interest**: Find the optimal (or near optimal) design to answer the question and use the design as a ‘template’ for sub-sampling the data.

2. **Sequential learning**: Apply a given design to incoming data or new datasets until the question of interest answered.

3. **Assess data quality**: Absence of design points/windows may indicate structured missingness or bias in the big dataset.

4. **Assess model quality**: Replicate designs can be ‘laid over’ the big data for model checking (eg posterior predictives), concept drift etc.

5. **Enlarge loss function**: Include model misspecification, time constraints etc.
Example: logistic regression

- 6 covariates
- 1,000,000 citizen-provided records

Analysis aims:
- Identification of important covariates for prediction
- Accurate and precise parameter estimates
Experimental design approach

• Select a random sample of 10,000 points to construct prior distributions.
• “Value add” to the information gained through a sequential design process.
• Use Sequential Monte Carlo for fast computation.

• For each new data point, update the prior information to reflect the information gained and form a 95% credible interval for all parameters.
• If any credible interval is contained within (−tol, tol), drop it from the model.
• Re-fit the reduced model and re-run.
• Iterate until 20,000 data points are extracted.
Example: Results

Designed approach:

<table>
<thead>
<tr>
<th>$tol$</th>
<th>Remaining covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>$x_1, x_2, x_3$</td>
</tr>
<tr>
<td>0.50</td>
<td>$x_2, x_3$</td>
</tr>
<tr>
<td>0.75</td>
<td>$x_3$</td>
</tr>
<tr>
<td>1.00</td>
<td>$x_3$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>SD</th>
<th>2.5th</th>
<th>Median</th>
<th>97.5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>-11.40</td>
<td>0.13</td>
<td>-11.67</td>
<td>-11.40</td>
<td>-11.16</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.42</td>
<td>0.03</td>
<td>-0.48</td>
<td>-0.42</td>
<td>-0.36</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.63</td>
<td>0.03</td>
<td>-0.68</td>
<td>-0.63</td>
<td>-0.56</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>3.03</td>
<td>0.05</td>
<td>2.94</td>
<td>3.03</td>
<td>3.13</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.20</td>
<td>0.03</td>
<td>0.12</td>
<td>0.20</td>
<td>0.26</td>
</tr>
</tbody>
</table>

✓ 3% of the full dataset analysed
✓ learn about “design holes”
✓ learn about model robustness
x 68 hours run-time (for each tol)

Comparison with random sampling:
✓ for each tol: 30 mins, i.e. ~ 136 such datasets in same time
x no random design provided as much information about the parameter values as that of the designed approach (across all values of tol).
CS and Bayesian modelling

Posterior = Likelihood / c * Prior

Meld → Design
Augment → Adjust
Elicit → History matching
3. Augmenting the likelihood

Citizen provides data $Z_i$ with bias $b_i$ and uncertainty (variance) $\sigma_i^2$

Model as measurement error:

ME model $Z_i \mid Y_i \sim N(Y_i + b_i, \sigma_i^2)$

Regression model $p(y \mid X, \beta)$

Prior model $p(\pi)$
Comparison of regressions with mixed data.
(a) shows the data, with the accurate (x) values marked by black crosses, and the less accurate (z) values marked by smaller grey crosses.
(b) shows the regression lines for the measurement error (ME) model, the validation alone analysis (validation) and the naïve analysis (naïve), as well as the true relationship (true).

Denham et al. (2011) The Bayesian conditional independence model for measurement error: applications in ecology
Comparison of naïve and ME models

Denham et al. (2011) The Bayesian conditional independence model for measurement error: applications in ecology
Case studies: citizen-provided observations

- Clear benefit from using the ME model, especially if we have a small amount of accurate data and access to a large amount of auxiliary data.

- Need for strict adherence to the assumptions made in the ME model. Even small errors in the assumptions in measurement error correction models can lead to even worse predictions.

- However, ignoring the citizen-data error can have serious results, especially if interest is in accurate estimation of all the parameters and the true underlying relationship between the predictors and the response.

Denham, Falk, M (2011) The Bayesian conditional independence model for measurement error: applications in ecology
CS and Bayesian modelling

Posterior = Likelihood / c * Prior

Meld
Design
Augment
Elicit
History matching
Adjust
4. Adjusting the likelihood

Weight the data via a power likelihood

\[ p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta)^{w_i} \]

\( w = \text{citizen ability} = f(\text{contributors, data sources}) \)

Fit \( w \) via an item-response model (or Rasch model)

\[ p(w) = c_i + \frac{1-c_i}{1+e^{-a_i(w-b_i)}} \]

\( a: \text{discrimination} \)
\( b: \text{difficulty} \)
\( c: \text{guessing} \)
Case study: monitoring the GBR with Peterson, Caley, Vercelloni, Bednarz, Brown et al.

reefvis.stats.technology

https://www.virtualreef.org.au/
Data sources

• **Professional sources:**
  o XL Catlin Seaview Survey (~20,000 geo-located images)
  o Great Barrier Reef Long-Term Monitoring Program (141 sites monitored regularly)
  o Reef Rescue Marine Monitoring Program (950 coral estimates)
  o the Heron Island survey (2000+ images taken in 2007 and 2012)
  o the Capricorn and Bunker group survey (video imagery)

• **Underwater images:**
  o extracted from videos collected by Reef Check Australia ([http://www.reefcheckaustralia.org](http://www.reefcheckaustralia.org))
  o annotated by citizen scientists

• **Covariates:**
  o Bathymetry, sea surface temperature, location, cyclones, bleaching, COTS

Each dataset provided multiple estimates of coral cover, but there were differences in the scale of the estimates and the estimation method.
Modelling citizen-derived data

- Image $j$, within survey $s$, annotated by citizen $p$:
  \[ w_{jps} = w_{ejs} w_{nps} w_{ap} \]

- Mean coral cover and weight for each cell $i$ in the image:
  \[ \bar{y}_{its} = \frac{\sum_{j \in J_{its}} w_{js} y_{js}}{\sum_{j \in J_{its}} w_{js}} \]
  \[ w_{its} = \sum_{j \in J_{its}} w_{js} \]

- Model:
  \[ \bar{y}_{its} = \text{Beta}(\mu_{it}, \phi) \]
  \[ \text{logit}(\mu_{it}) = X_{it} \beta + u_i + v_t \]
Results: northern region, multiple years
Results: temporal trend

Inner reef

Outer reef
Accuracy of citizen-derived data

Compared to a marine scientist (with accuracy = 1).
Black dots: a user’s mean accuracy across all images they annotated
Width of the boxplot: proportional to the number of annotations performed
How to choose the weights?

Structure determined by 3 eminent experts

Caley et al. (2014) What is an expert? A systems perspective on expertise
What makes an expert?

Weights determined by 4 supra-experts

- Quality of work
- Taxonomic descriptions
- Total productivity
- Geographic reach
What makes an expert?

Results based on Bayesian Network model for expertise
CS and Bayesian modelling

Posterior = Likelihood / c * Prior

Meld → Design
Augment
Elicit → History matching
Adjust
5. CS + history matching

\[
\text{logit}(p_i) = \beta_0 + \beta_1 x_i
\]

Priors: \(\beta_0 \sim \text{Cauchy}(0, \lambda_1)\)

\(\beta_1 \sim \text{Cauchy}(0, \lambda_2)\)

Consider the \textit{prior predictive distribution}

\[
p(y) = \int p(\theta)p(y|\theta)d(\theta)
\]

and use “history matching”

Wang, Nott, Drovandi, M, Evans (2016)
History matching priors

1. Choose a set of summary statistics $S^j$, based on data to be observed with density $p(y|\theta)$.

2. For each statistic, specify the set of values $S^j_0$ that would be considered surprising if they were observed.

3. Let $p(S^j|\lambda)$ be the prior predictive distribution for $S^j$.

4. Compute $p_j(\lambda) = P(\log p(S^j|\lambda) \geq \log(p(S^j_0|\lambda)))$.

5. Make a decision based on a threshold chosen according to a “degree of surprise”.

Use ABC

Repeat in waves, to obtain non-implausible priors.
Example: logistic regression

• **Summary statistics?**

  sum of variances of responses

• **Surprising?**

  If all $\hat{p}_i$ are equal to either 0.01 or 0.99 (so $S_0^1 = 0.198$)

• **Not surprising?**

  If the $\hat{p}_i$ are “smooth”
  
  e.g. $\hat{p}_1 = 0.01$, $\hat{p}_2 = 0.25$, $\hat{p}_3 = 0.75$, $\hat{p}_4 = 0.99$
  
  (so $S_0^2 = 1.974$).
Example: logistic regression

For $S^1=0.198$ and $S^2=1.974$, compute predictive p-values for the summary statistics over a grid of 10,000 $\lambda$ values.

Plot of conflict p-values for $S^1$ (left, light blue is good) and $S^2$ (right, pink is good)
CS and Bayesian modelling

Posteriors = Likelihood / c * Prior

Meld
Design
Augment
Adjust
Elicit
History matching
Option 1: Eliciting priors on relationships

\[
\Pr(Y_i=1 \mid X) = \Sigma \beta_j X_{ij} + e_i
\]

\[
\beta_j \sim N(B_j, \sigma_j^2)
\]

“Given the other variables in the model, what is the impact of \(X_j\) on probability of presence, and how certain are you of that?”
Option 2: Eliciting priors on outcomes

\[ Y_i \sim \text{Beta}(\mu_i, \sigma_i) \]

\[ \text{logit}(\mu_i) = \sum_{j=1}^{p} \beta_j X_{ij} + e_i \]

\[ \tilde{\beta}_j \sim N(0, \tau_j) \]

Use *induced* priors on coefficients

\[ \beta_j | Y, X \sim N(\tilde{\beta}_j, g\tau_j) \]
Combining responses

Mixture model
• each citizen has their own distribution

Hierarchical model
• account for various sources of variation (e.g. within and between experts)
• account for potential dependence between experts

Albert et al. (2012) Combining Expert Opinions in Prior Elicitation
Case Study:
Bayesian analysis of complex systems

Wu, Caley, et al.

Can we find “ecological windows” for dredging to reduce the impact on seagrass?
### Table: G

<table>
<thead>
<tr>
<th>E</th>
<th>F</th>
<th>normal</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>low</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>medium</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>no</td>
<td>low</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>medium</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>
• Model cumulative impacts of dredging scenarios

• Combine data from many sources, for 28 seagrass & coastal development sites worldwide

• Three criteria:
  o Resistance (<20% loss of species)
  o Recovery (to within 20%)
  o Persistence (no increase in risk of local extinction)

*With Paul Wu et al.*
• 36 node Bayesian Network to encode conditional probabilistic relationships & interactions between system factors
• DBN to capture cumulative effects and feedback processes
• Non-homogeneous DBN to model the multiple system transition rules for decline & recovery
World distribution of seagrasses (green) & ports (blue dots).
Heat maps: average recovery time (bottom panel) & average ratio of extinction risk to baseline risk
Bars correspond to dredging periods: 1, 2, 3, 6, 9, 12 mths
Labels coloured by genera – *Halophila*, *Zostera*, *Amphibolis*. 
• Seagrass population resilience responses for 3 sites (rows) & 8 dredging scenarios & a control scenario (cols).
• Each pie slice shows time to recovery for dredging starting in that month.
• Months are given by numerals in the outer ring where 12 denotes December.
• Outer edge of each pie: resilience criteria score, from dark green (all criteria satisfied), green, orange, yellow (loss but recovery within 6 months), to red for no criteria satisfied.
Results

For seagrass meadows subjected to dredging stress, ecological windows can maximise resilience globally, achieving up to fourfold reduction in recovery time and 35% reduction in extinction risk.
Case Study

Creating a jaguar corridor across the Peruvian Amazon
Research Team: “Eyes on the Wild”

Tomasz Bednarz
June Kim
Ross Brown
Kevin Burrage
Jacqueline Davis
Claudia Deasy
Vanessa Hunter
Alan James
Efrom Lloyd
Kerrie Mengersen
Erin Peterson
Steve Psaltis
Julie Vercelloni
Gavin Winter
Nan Ye

Media
Maths
Stats
Machine learning
Visualisation
Project Aims

• To create a “Peruvian Amazon Conservation Corridor” with the jaguar as the key focus

• To merge VR (and other new tech such as drones, bioacoustics, camera traps, smart phones) with statistical modelling

• To better engage local and international experts in conservation
Two onsite locations:

1. Imiria Reserve
2. Pacaya Samiria Reserve
Imiria Reserve
Pacaya Samiria Reserve
Primary information sources

Data
Primary information sources

Local people
Primary information sources

Experts
Eliciting information: in Peru

“Tell us your stories”
Eliciting information: in Peru

“Stars on a map”
New data
New reality

Go-pros
360 cameras
3D cameras
VR headset
+ GPS
“Putting the expert into the environment”
Maths & Stats Models

• Spatial model of jaguar encounters

• Markov model of jaguar movements

• Bayesian network of key factors and impacts

• Bayesian models using data
“So what?”
Conclusions

• Citizen-provided data and analysis is a cost-effective resource with other potential benefits.

• Concerns re accuracy and variability can be reduced by:
  o Collecting reference data and analyses for calibration
  o Close supervision
  o Qualitative and quantitative quality assurance checks
  o Designing tasks with the skill of the citizens in mind
  o Training
  o Correcting for potential bias and uncertainty in the modelling.