QUT Centre for Data Science

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Bayesian Modelling and Analysis of Challenging Data

Kerrie Mengersen
School of Mathematical Sciences
QUT

PC Mahalanobis Lecture Series
January 2021

Programme of Lectures

January 27th:

- Lecture 1: 10-1045am IST (230pm-3:15pm AEST)
  Identifying the Intrinsic Dimension of High-Dimensional Data

- Lecture 2: 11-11:45am IST (3:30pm-4:15pm AEST)
  Finding Patterns in Highly Structured Spatio-Temporal Data

January 29th:

- Lecture 3: 10-1045am IST (230pm-3:15pm AEST)
  Describing Systems of Data

- Lecture 4: 11-11:45am IST (3:30pm-4:15pm AEST)
  Making New Sources of Data Trustworthy

Bayesian Modelling and Analysis of Challenging Data

Lecture 4:
Making new sources of data trustworthy

Cath Leigh, Erin Peterson,
Edgar Santos-Fernandez, Julie Vercelloni

Case study 1: Using citizen science data to inform models

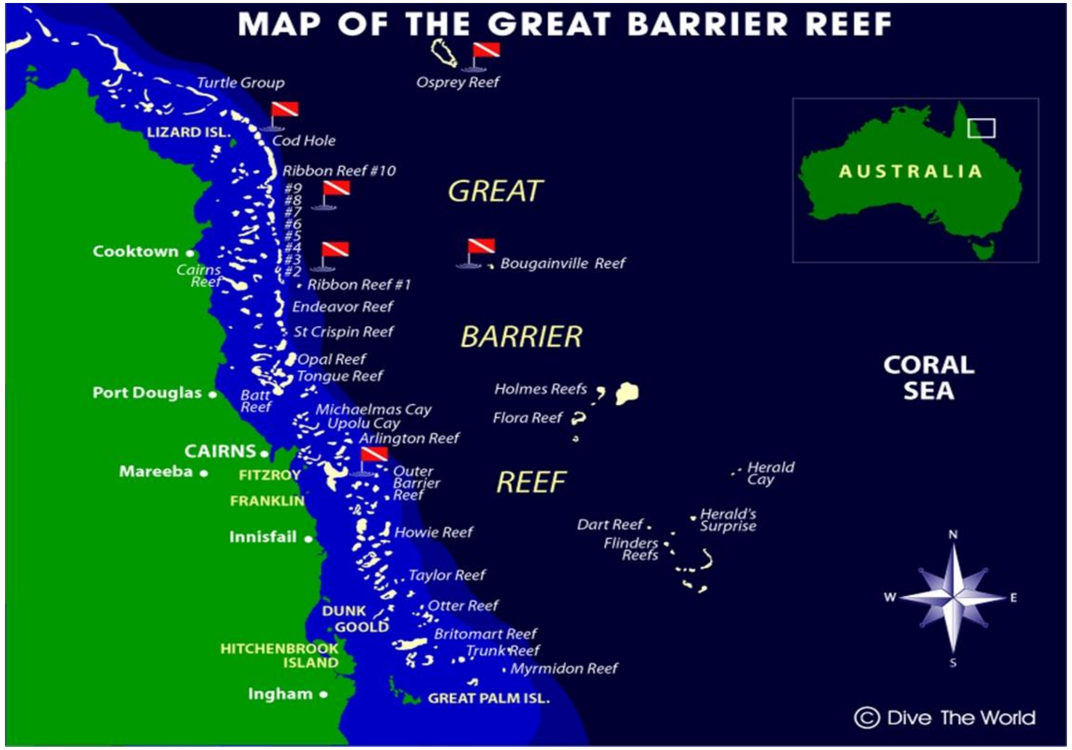
QUT Queensland University of Technology

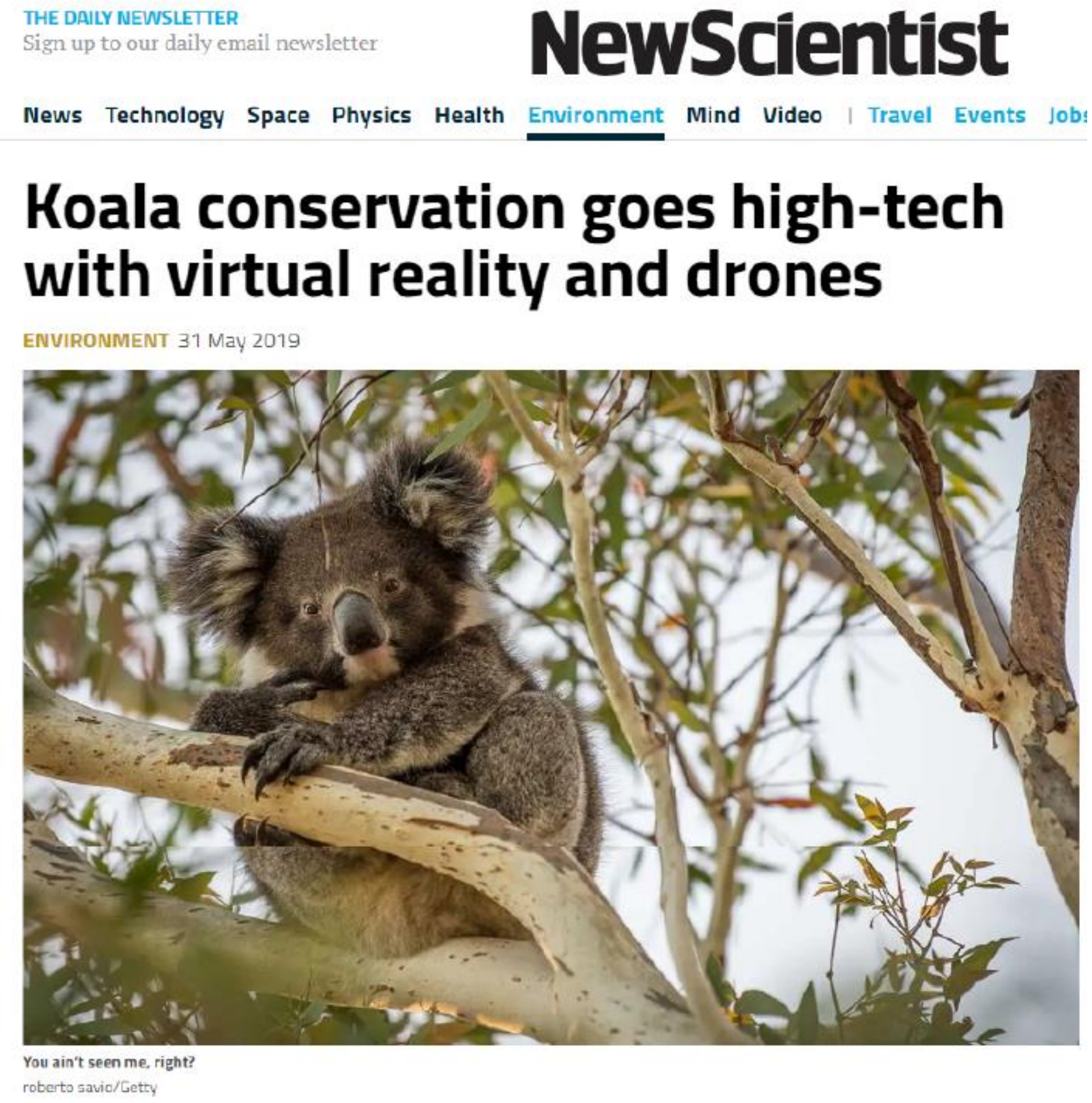
crc•si

Australian Institute of Marine Science

ACEMS Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers

Queensland Government





Koala conservation goes high-tech with virtual reality and drones

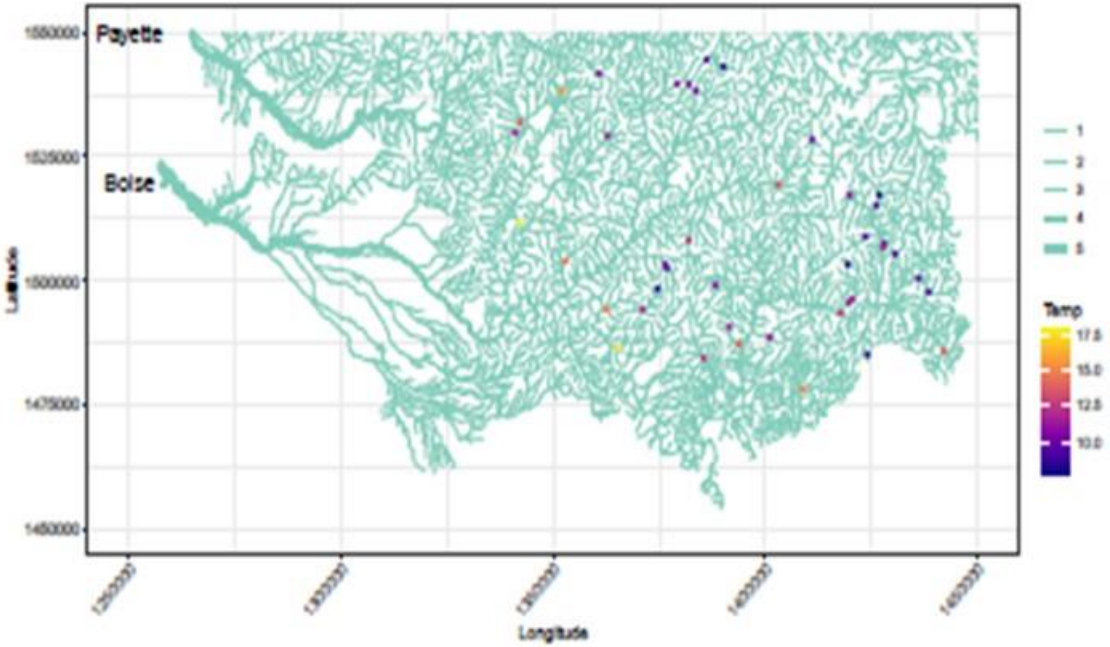
ENVIRONMENT 31 May 2019

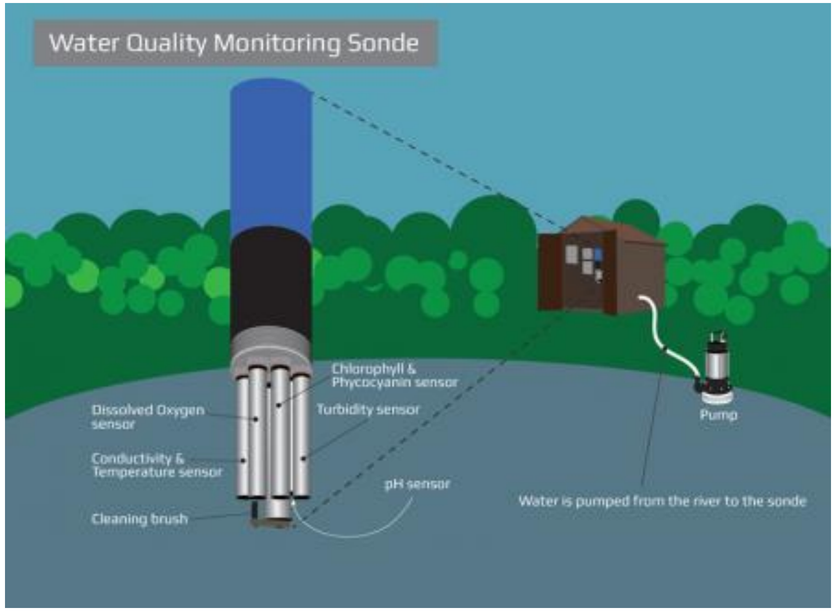
You ain't seen me, right?

roberto savio/Getty

C. Leigh *et al.* (2019) Using virtual reality and thermal imagery to improve statistical modelling of vulnerable and protected species. *PLoS ONE.*

Case study 2: Using low-cost sensor data to inform models

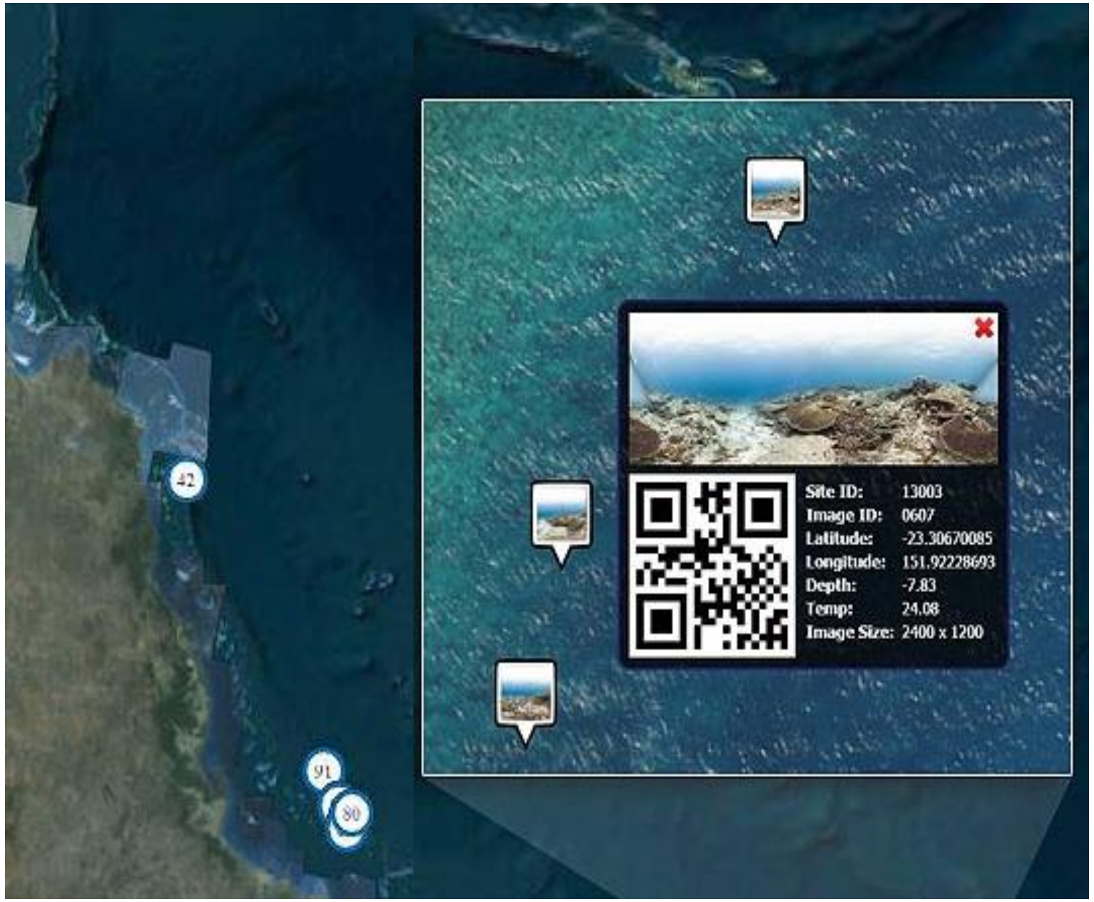



Water Quality Monitoring Sonde
Chlorophyll & Phycocyanin sensor
Turbidity sensor
Dissolved Oxygen sensor
Conductivity & Temperature sensor
pH sensor
Cleaning brush
Pump
Water is pumped from the river to the sonde

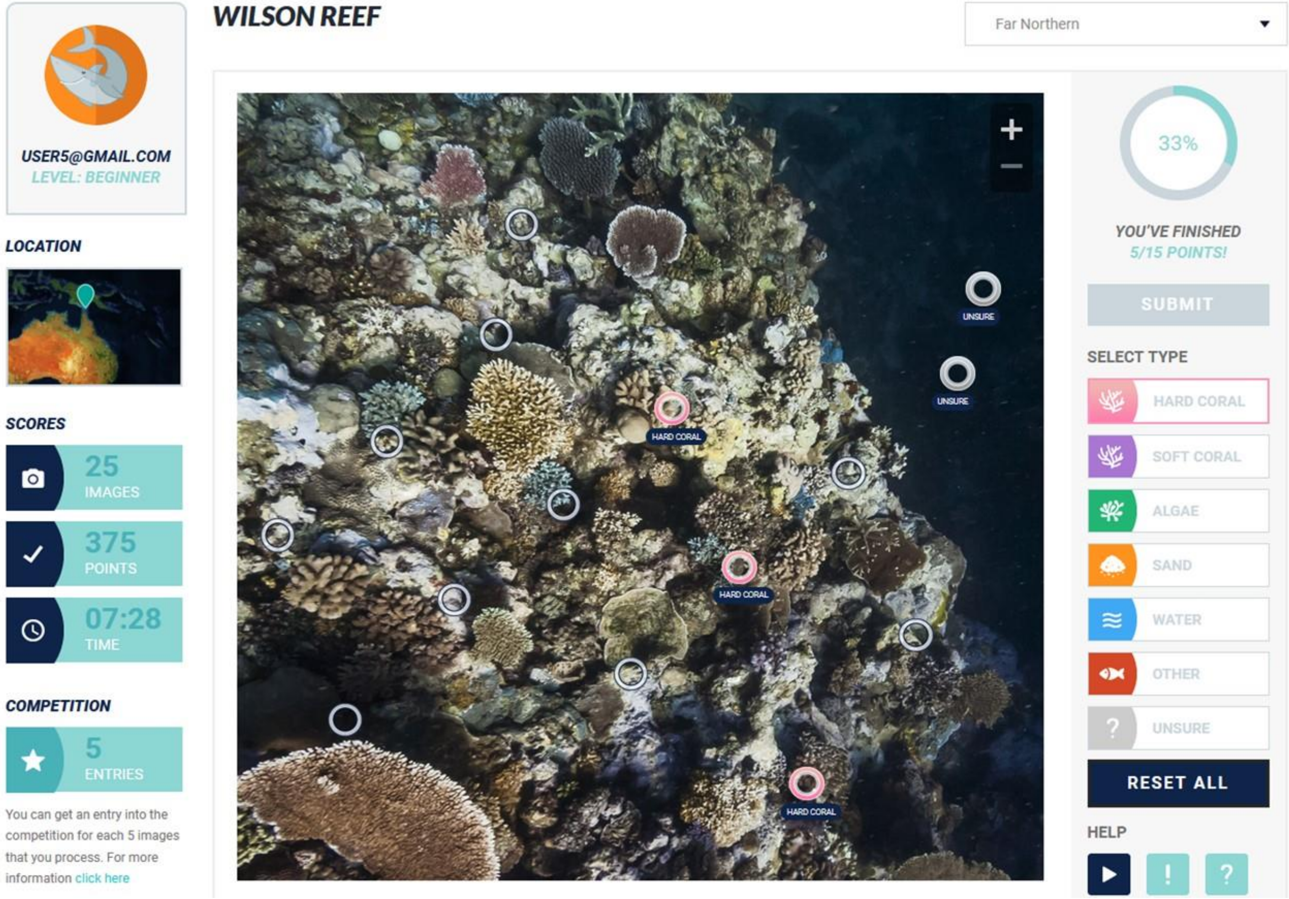
Source: US EPA

Case Study 3: Explainable models

Case study 1: Using citizen science data to inform models

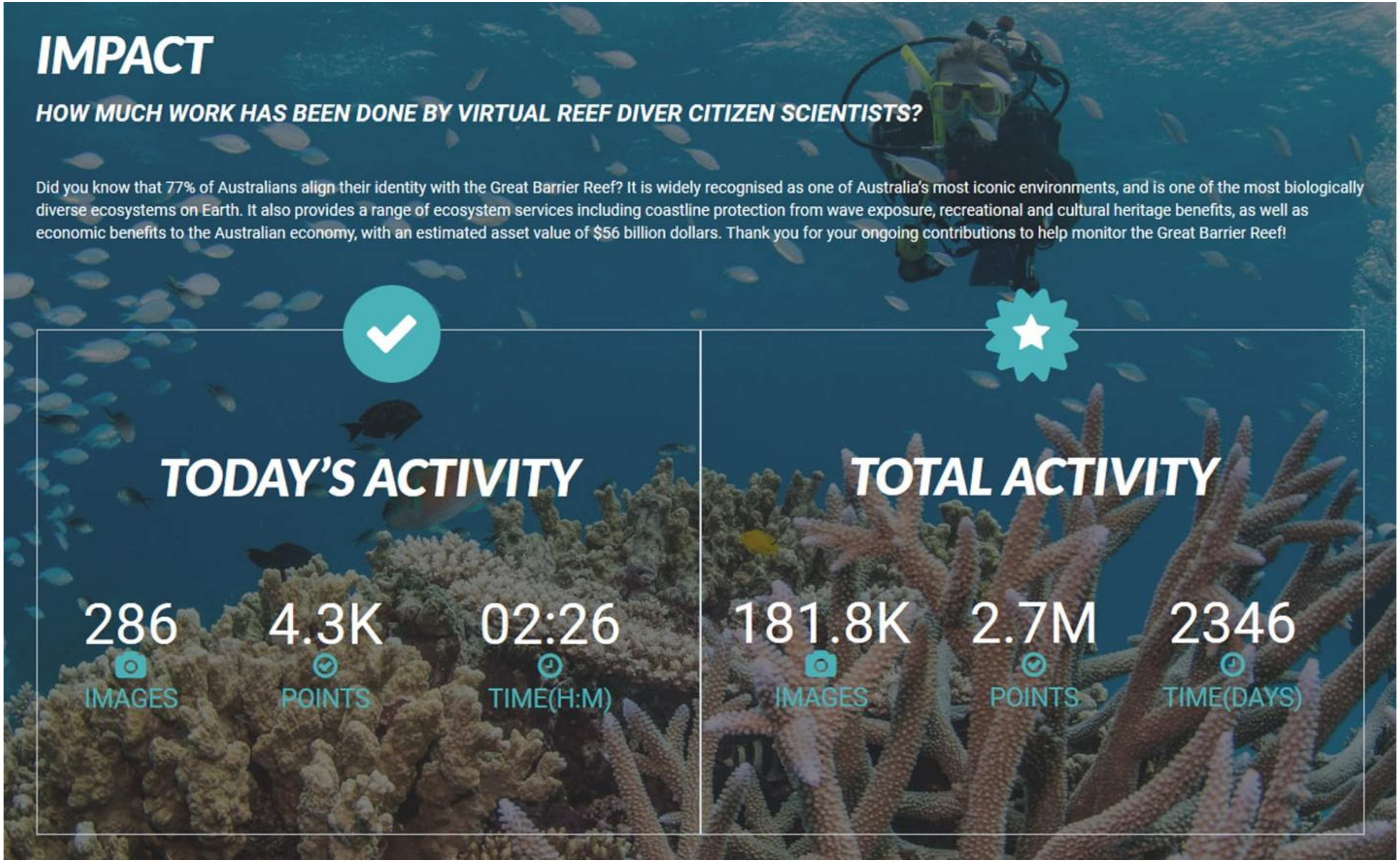


Site ID: 13003
Image ID: 0607
Latitude: -23.30670085
Longitude: 151.92228693
Depth: -7.83
Temp: 24.08
Image Size: 2400 x 1200



You can get an entry into the competition for each 5 images that you process. For more information click here

YOU'VE FINISHED 5/15 POINTS!



IMPACT

HOW MUCH WORK HAS BEEN DONE BY VIRTUAL REEF DIVER CITIZEN SCIENTISTS?

Did you know that 77% of Australians align their identity with the Great Barrier Reef? It is widely recognised as one of Australia's most iconic environments, and is one of the most biologically diverse ecosystems on Earth. It also provides a range of ecosystem services including coastline protection from wave exposure, recreational and cultural heritage benefits, as well as economic benefits to the Australian economy, with an estimated asset value of $56 billion dollars. Thank you for your ongoing contributions to help monitor the Great Barrier Reef!

TODAY'S ACTIVITY

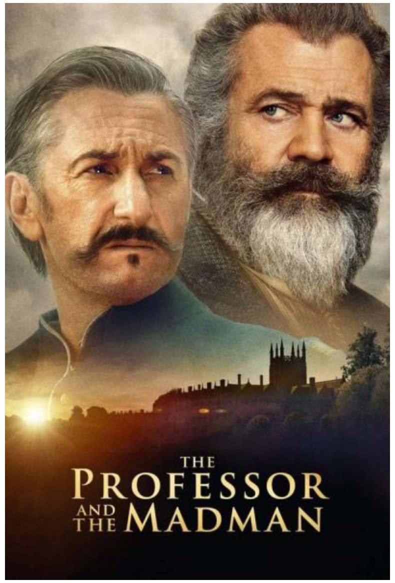
286	4.3K	02:26
IMAGES	POINTS	TIME(H:M)

TOTAL ACTIVITY

181.8K	2.7M	2346
IMAGES	POINTS	TIME(DAYS)

Analysing crowdsourced data

“the data source of the 21st century”



- Participatory method of building or analysing a dataset
- Many platforms, eg mobile apps, internet marketplaces (Amazon Mechanical Turk)

- Advantages: Cheap, real-time, numerous, widespread observations

- Tradeoff: precision, quality, sampling frame, simple tasks/questions

Crowdsourcing in ecology

- Popularity:
    - Hundreds of CS projects
    - e.g. Federal Crowdsourcing and Citizen Science Toolkit (US Federal, 2018), Zooniverse, iNaturalist, eBird

- Concerns:
    - inherent presence of misclassification or measurement errors resulting from participants' variable skill levels and abilities.
    - spatial dependence in the data must be taken into account.
    - citizen scientists tend to capture observations in easily accessible areas.

Approaches to modelling error-prone CS data

1. Ignore measurement error.

2. Weighted linear regression with observation weights proportional to the user's accuracy or performance measures (Peterson et al., 2020).

3. Take into consideration the user's sensitivity (se) and specificity (sp): ability to correctly detect presence/absence (true positive/negative) of target species.

spatially dependent misclassification error (SDME) approach

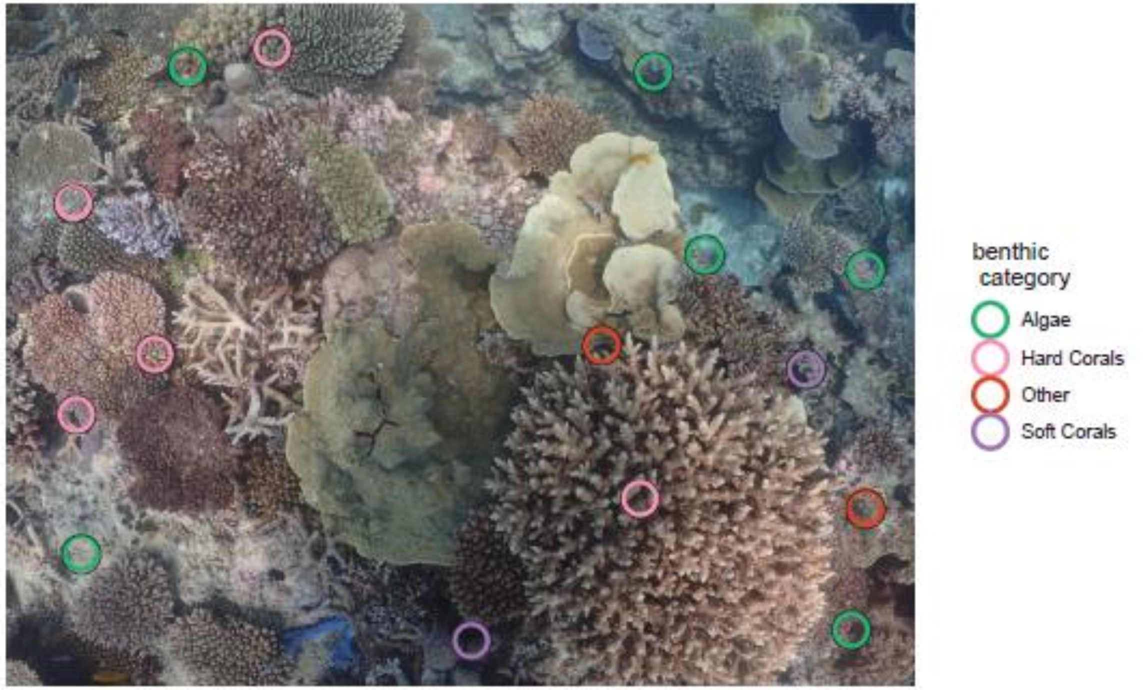
Case study

- Experiment within the Virtual Reef Diver project.
- Aims:
  - determine impact of different reef disturbances on hard coral cover changes
  - assess participants' abilities to identify hard corals within geotagged images
  - evaluate the quality of the estimates obtained from the experiment.
- Four covariates: Degree heating weeks (max DHW/yr), no-take (binary, marine reserve), outer (outer/inner reef), cyc (cumulative hrs exposure to waves in cyclone)
- N = 1585 images from unique locations in GBR, 2008-2017
- Each image had 40 spatially balanced, random points, classified by experts.

Want to expand experts' data to larger regions → CS.

Crowdsourcing via Mechanical Turk

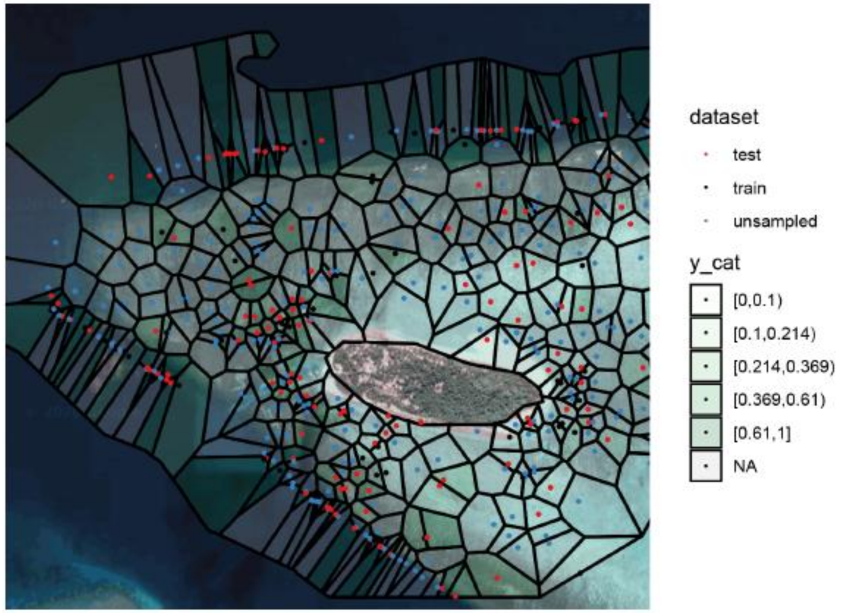
- Displayed 514 images on Amazon Mechanical Turk
- Workers asked to classify points into 5 benthic categories (hard coral, soft coral, algae, sand, other)
- Help file, qualification test (>60% to pass)
- Data collected Jan-14 to Feb-12 2020.
- Participants paid 0.10 USD per completed image (> U.S. federal min. wage $7.25/hr)
- Assigned up to 40 random images, classify 15 points per image.



Setup

- Spatial model using stochastic partial differential equations (SPDE) via STAN
- Images defined by latitude & longitude, centroids of Voronoi polygons
- Presence/absence of target class (hard coral) in a subset of points (*q* = 15) within an image:
    - q: No. points within an image ($q = 15$)
    - y: True proportion of points containing target class
    - $z_{ijk} = (0,1)$ Presence/absence of target class in point k from image j classified by subject i
    - $\hat{y}_{ij}$: apparent proportion from CS

$$\hat{y}_{ij} = \sum_{k=1}^{q} z_{ijk} / q$$



Voronoi diagram: Heron Island, GBR
Test & training data
Coral cover (y_cat) classified by experts

Performance of subjects

$$se_i = \frac{\sum_{j=1}^{m} \sum_{k=1}^{q} TP_{ijk}}{\sum_{j=1}^{m} \sum_{k=1}^{q} TP_{ijk} + \sum_{j=1}^{m} \sum_{k=1}^{q} FN_{ijk}}$$

$$sp_i = \frac{\sum_{j=1}^{m} \sum_{k=1}^{q} TN_{ijk}}{\sum_{j=1}^{m} \sum_{k=1}^{q} TN_{ijk} + \sum_{j=1}^{m} \sum_{k=1}^{q} FP_{ijk}}$$

$$acc_i = \frac{\sum_{j=1}^{m} \sum_{k=1}^{q} TP_{ijk} + \sum_{j=1}^{m} \sum_{k=1}^{q} TN_{ijk}}{\sum_{j=1}^{m} \sum_{k=1}^{q} TP_{ijk} + \sum_{j=1}^{m} \sum_{k=1}^{q} FN_{ijk} + \sum_{j=1}^{m} \sum_{k=1}^{q} TN_{ijk} + \sum_{j=1}^{m} \sum_{k=1}^{q} FP_{ijk}}$$

$$\hat{y}_{ij} = y_j \times se_i + (1 - y_j) \times (1 - sp_i)$$

Model

$$y_j|\alpha_{j,}\beta_j \sim \mathrm{Beta}(\alpha_{j,}\beta_j)$$

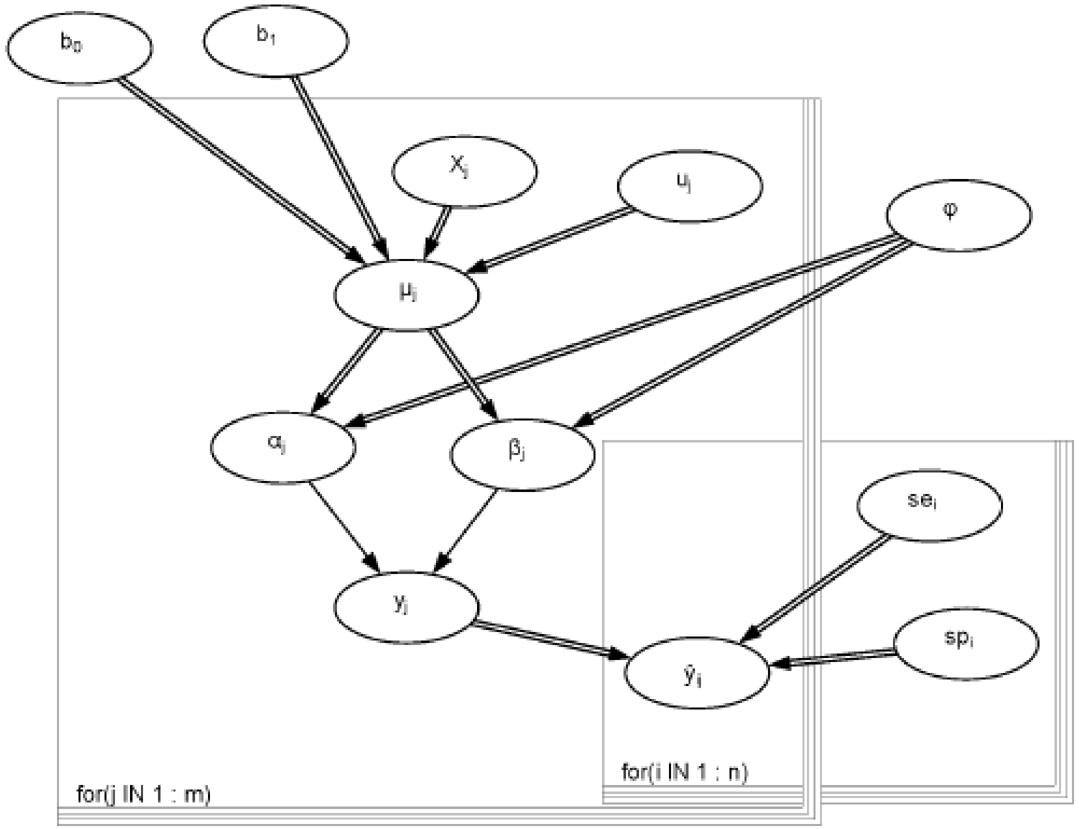
$$\alpha_j = \mu_j\phi\,;\ \beta_j = -\mu_j\phi + \phi$$

$$\mu_j = \mathrm{E}\left(y_j|\alpha_{j,}\beta_j\right)$$

$$Var\left(y_j\right) = \mu_j(1-\mu_j)/(1+\phi)$$

$$\mathrm{logit}\left(\mu_j\right) = X_j b + u_j + \varepsilon_j$$

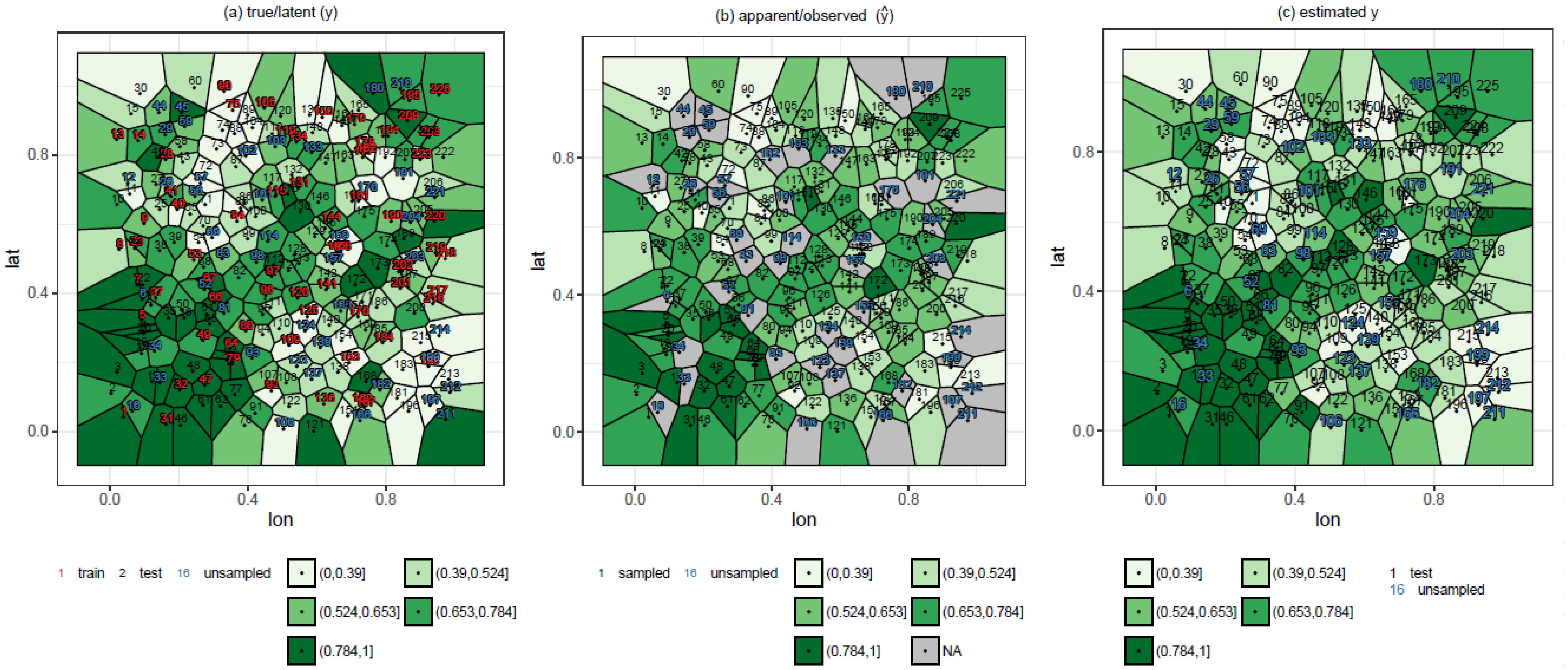
$$u_l|u_t, \tau_u \sim \mathrm{N}\left(\frac{1}{n_l}\sum_{l\sim t} u_t\,, \frac{1}{\tau_u n_l}\right)$$



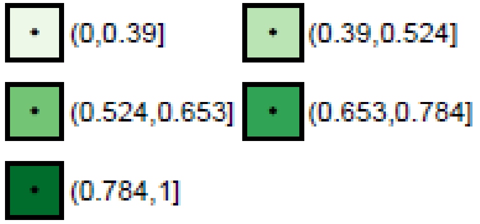
- Hamiltonian Monte Carlo (HMC) simulation in Stan, based on the no-U-turn sampler (NUTS).
- 3 chains each with 60,000 samples, half burn-in, thin 1 in 3.

225 locations
67 known

Simulation study

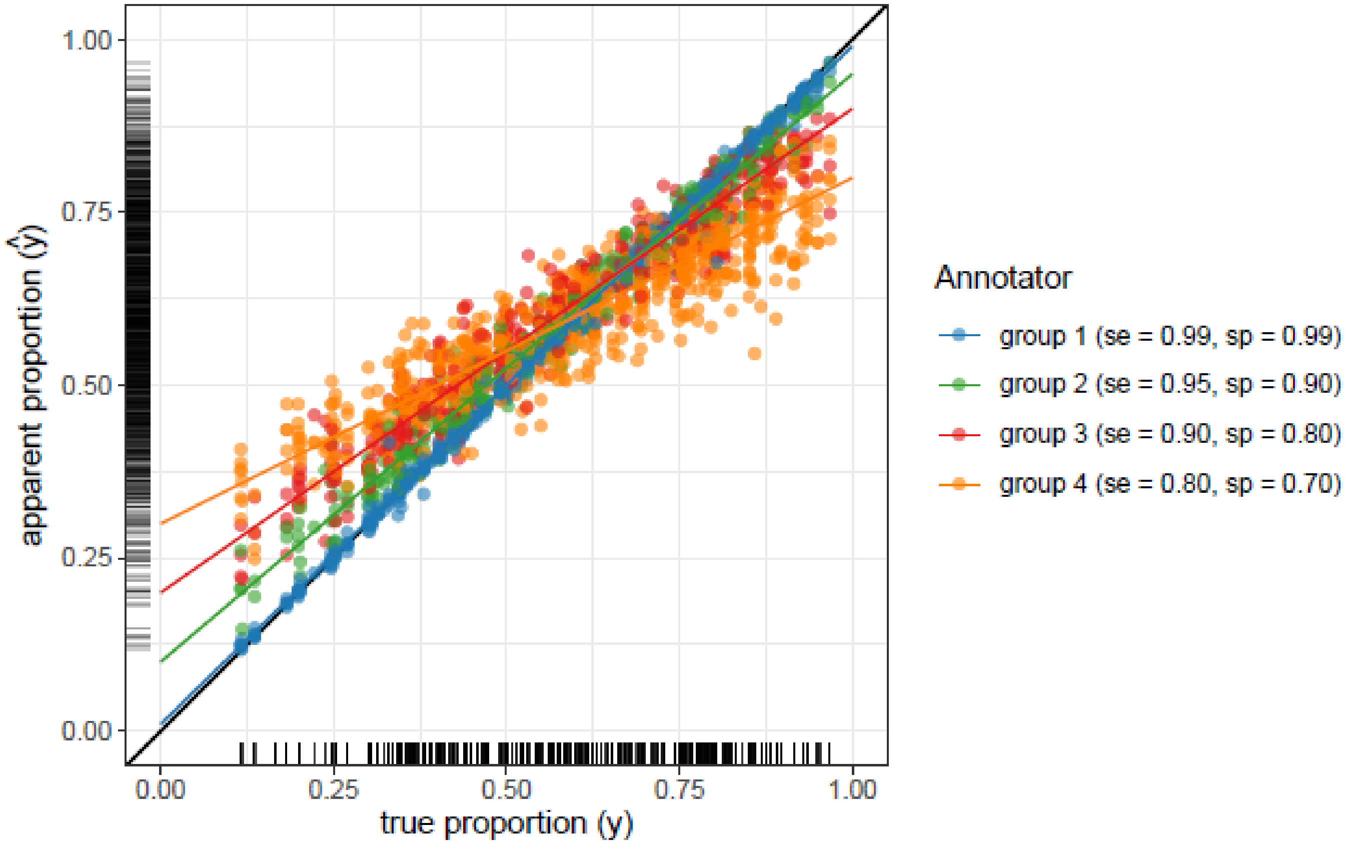


Test locations Training locations Unsampled locations

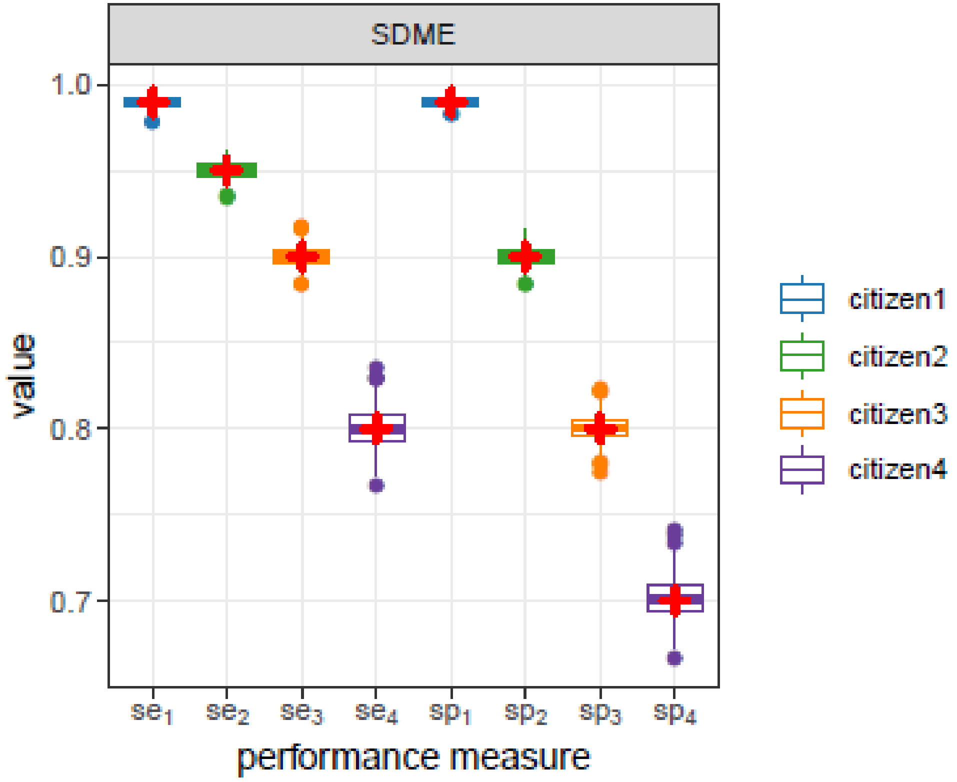


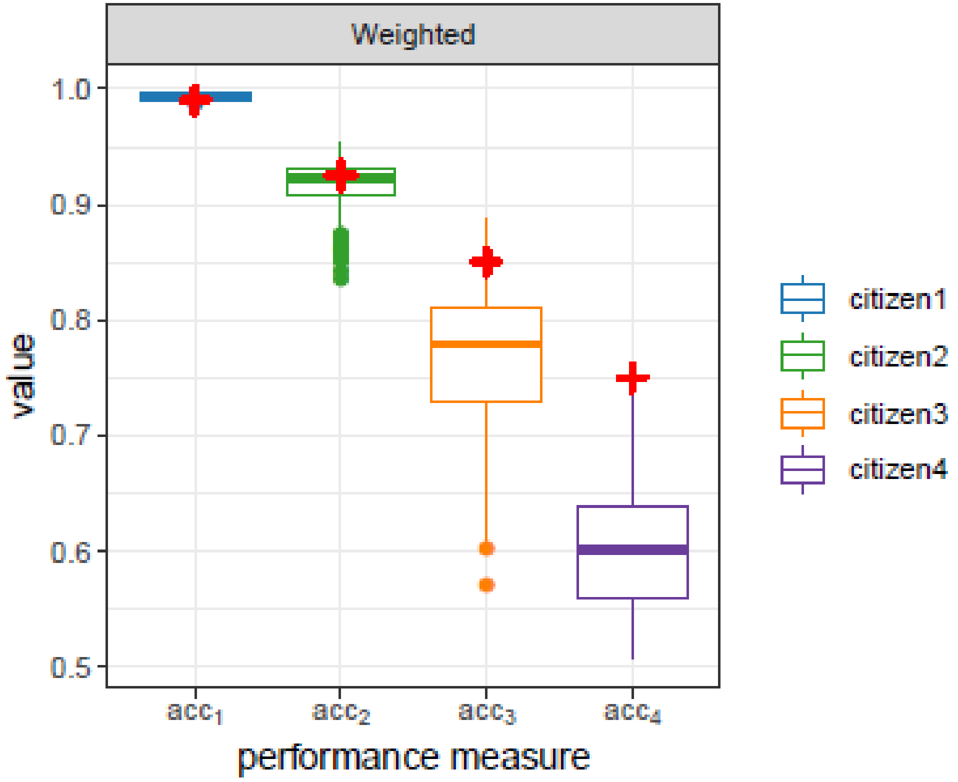
Results – classes of user ability

- About half responses slightly over-estimated.
- Exact class obtained in 199 out of 225 locations.
- The SDME model captured the true parameter values much better than the weighted model.



Results – classes of user ability

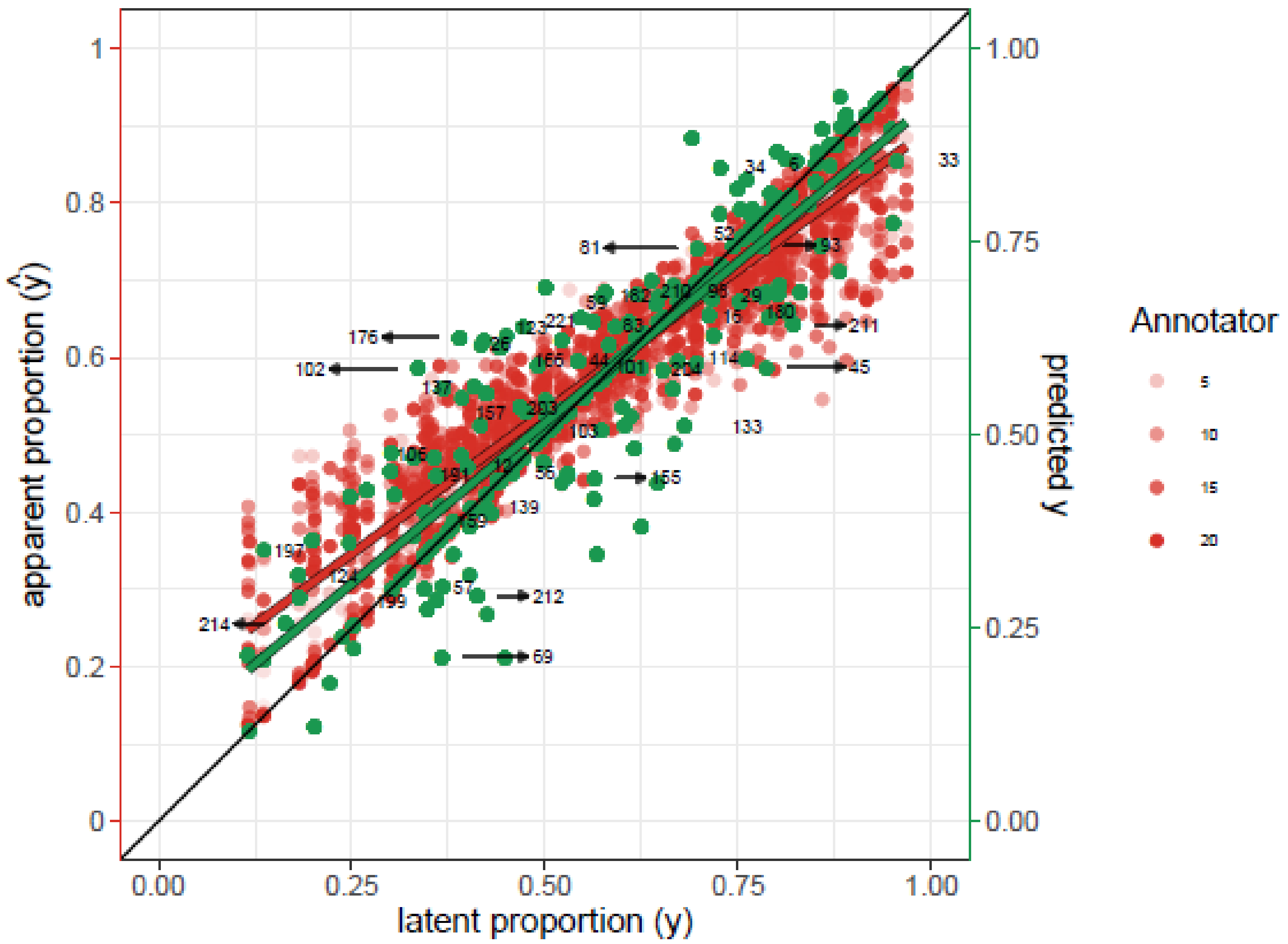




Results – bias correction from SDME model

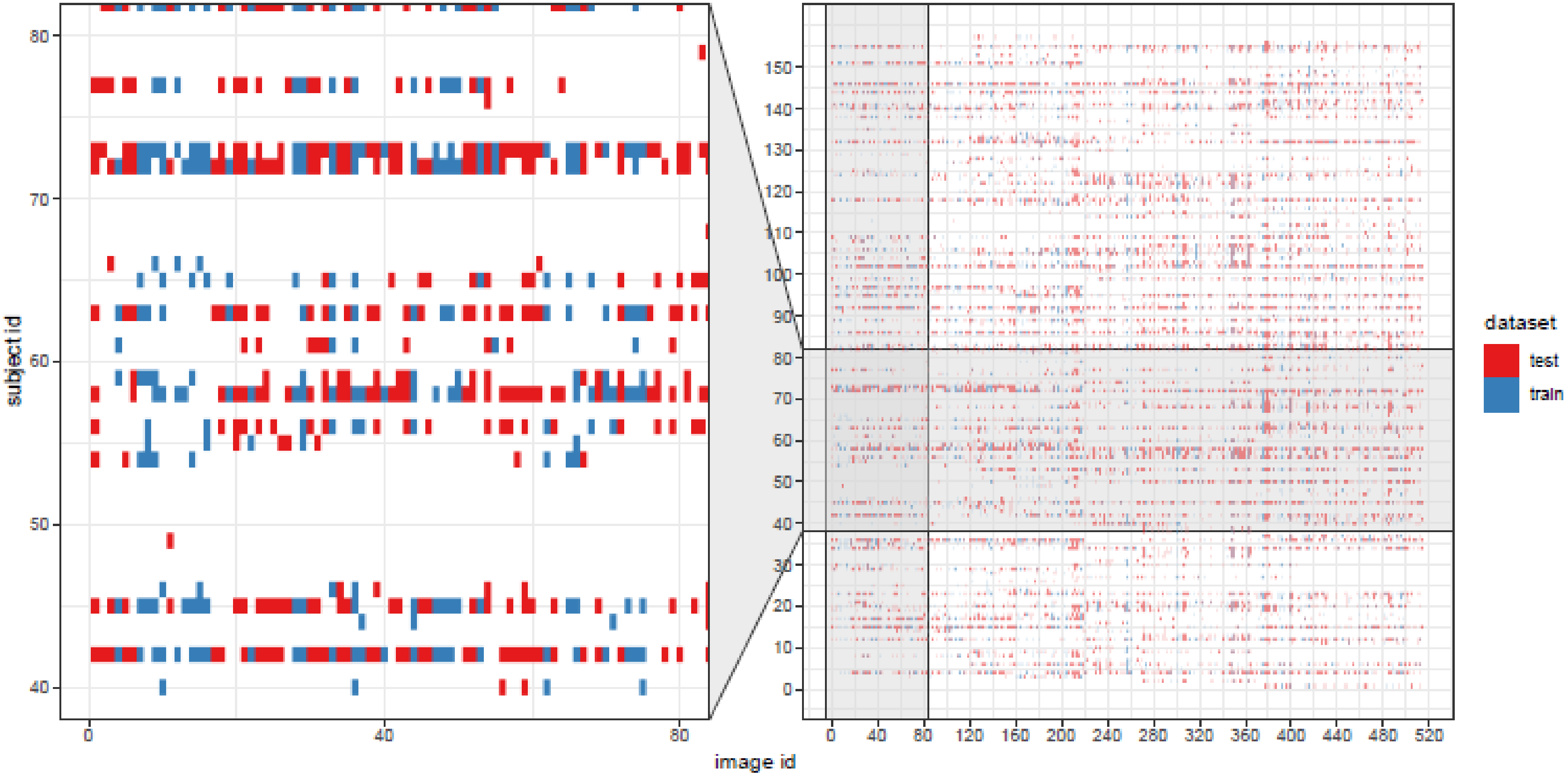
O elicited $\hat{y}$

O estimated latent variable y_{estim} after accounting for misclassification errors



Case Study Results

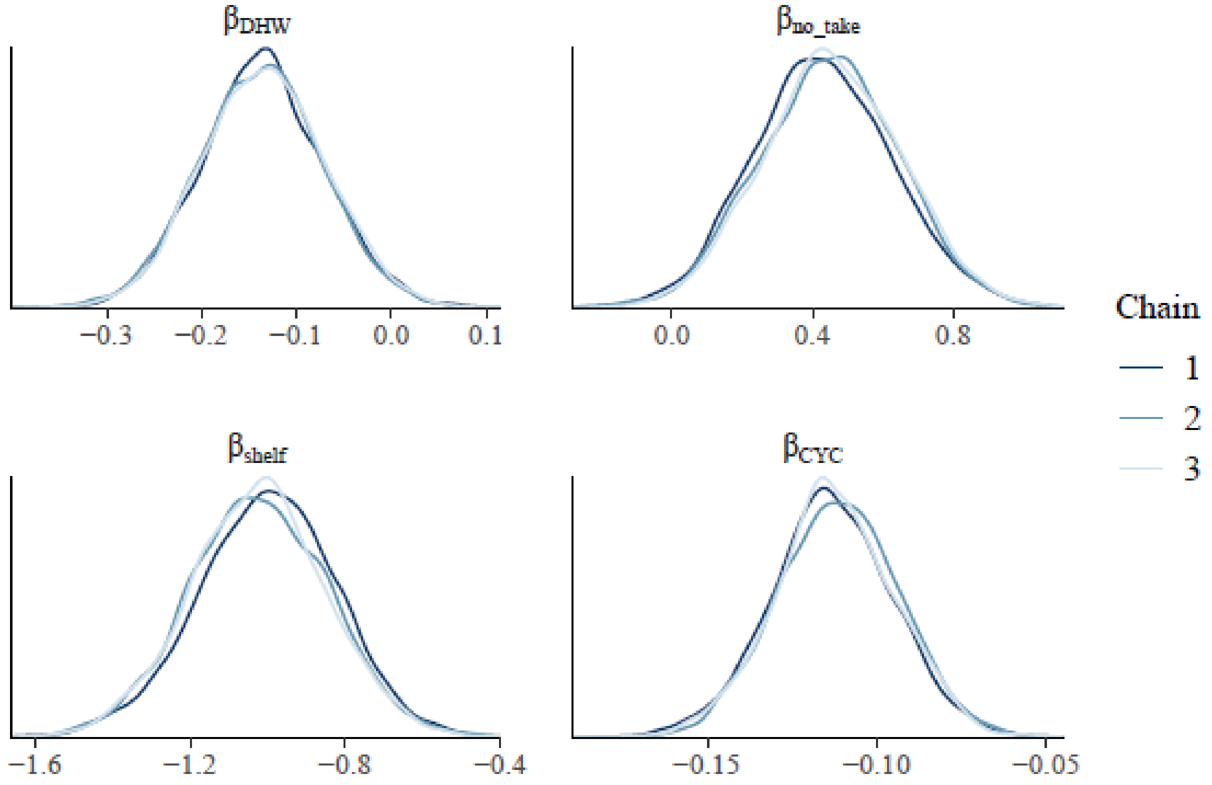
Images classified per user



Case Study Results

- DHW and cyclone impact covariates have a substantial negative effect on the proportion of hard corals (97.5% CrI << 0).

- No-take marine reserves and middle shelf reefs tend to have substantially higher proportions of corals.



Alternative IRT approach to correcting CS data

- Aim to account for user ability and task difficulty in a geospatial setting

- Build on Rasch model (Rasch, 1960) → 3 parameter logistic model (3PL) → GLM perspective → allow for space and time → dynamic response to map latent variables over time → dynamic IRT for sequential updates

- Propose 3PLUS model : extension of 3PL (U)sing (S)patially dependent item difficulties and variation of the linear logistic test model (LLTM)

3PLUS model

For user i, image j, point k, ask if the point contains the target class (Yes/No)

$$Y_{ijk} \sim \mathrm{Bern}\left(p_{ijk}\right)$$
$$p_{ijk} = \eta_j + (1 - \eta_j)\frac{1}{1 + \exp\{-\alpha_j(\theta_i - b_j)\}}$$

- θ_i : latent ability of the ith user
- b_j : difficulty of the jth image
- α_j : discrimination parameter
- η_j : pseudoguessing parameter

Can extend to include Covariates associated with ability and difficulty

$$b_j | b_m, \tau_b \sim \mathrm{N}\left(\frac{1}{n_l}\sum_{l\sim m} b_m\,, \frac{1}{\tau_b n_l}\right)$$

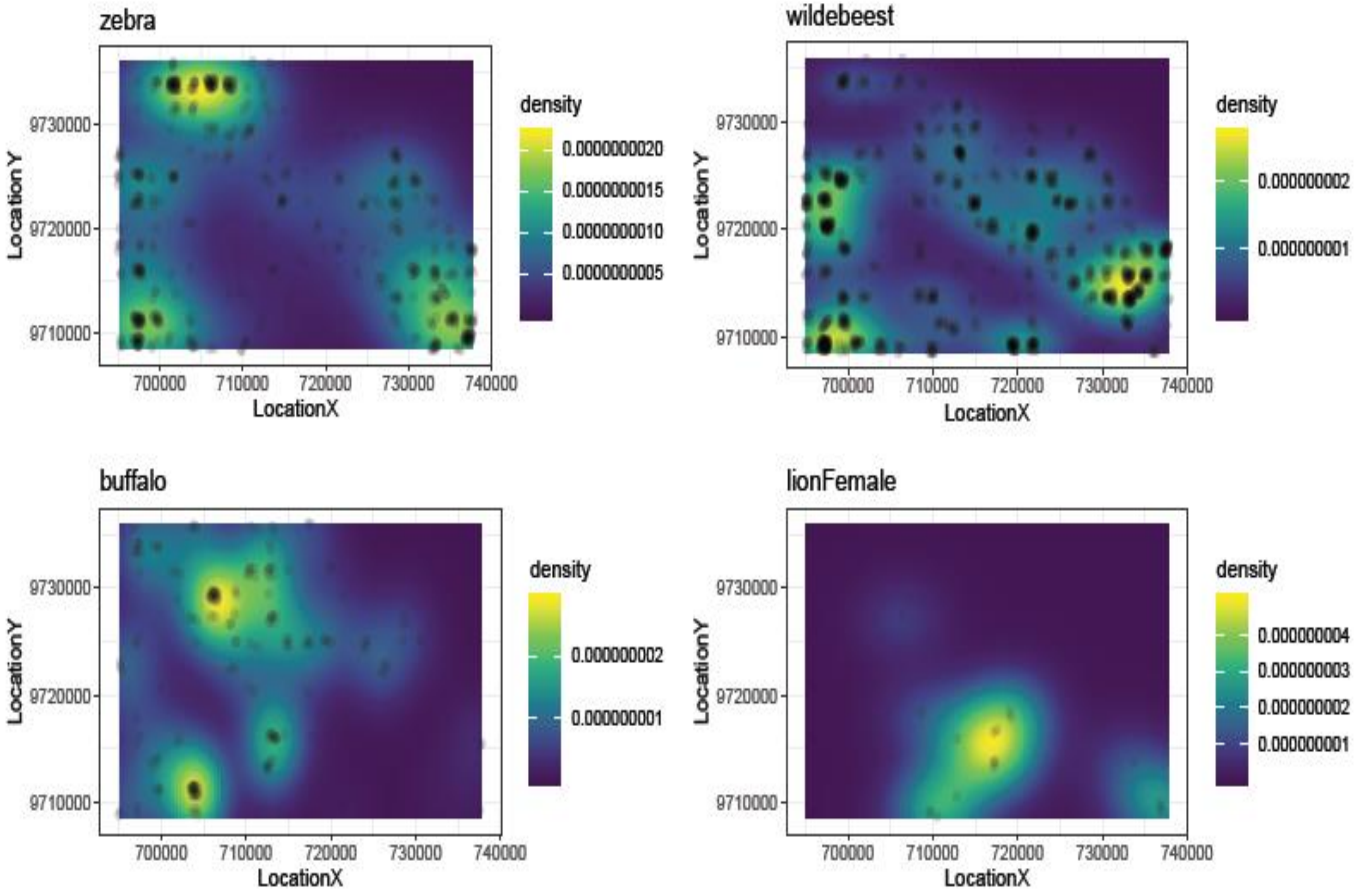
Simulation study

- Same study as before

- 3PLUS model vs 3PL model:
  - improved prediction accuracy (80% vs 62.2%)
  - smaller RMSE (0.26 vs 0.48), WAIC (44587 vs 44637), LOO (44588 vs 44639)
  - improved correlations between estimated and true difficulties (0.97 vs 0.88)
  - similar estimates of discrimination, but better estimate of pseudoguessing

Case study – *“Hakuna my data”*

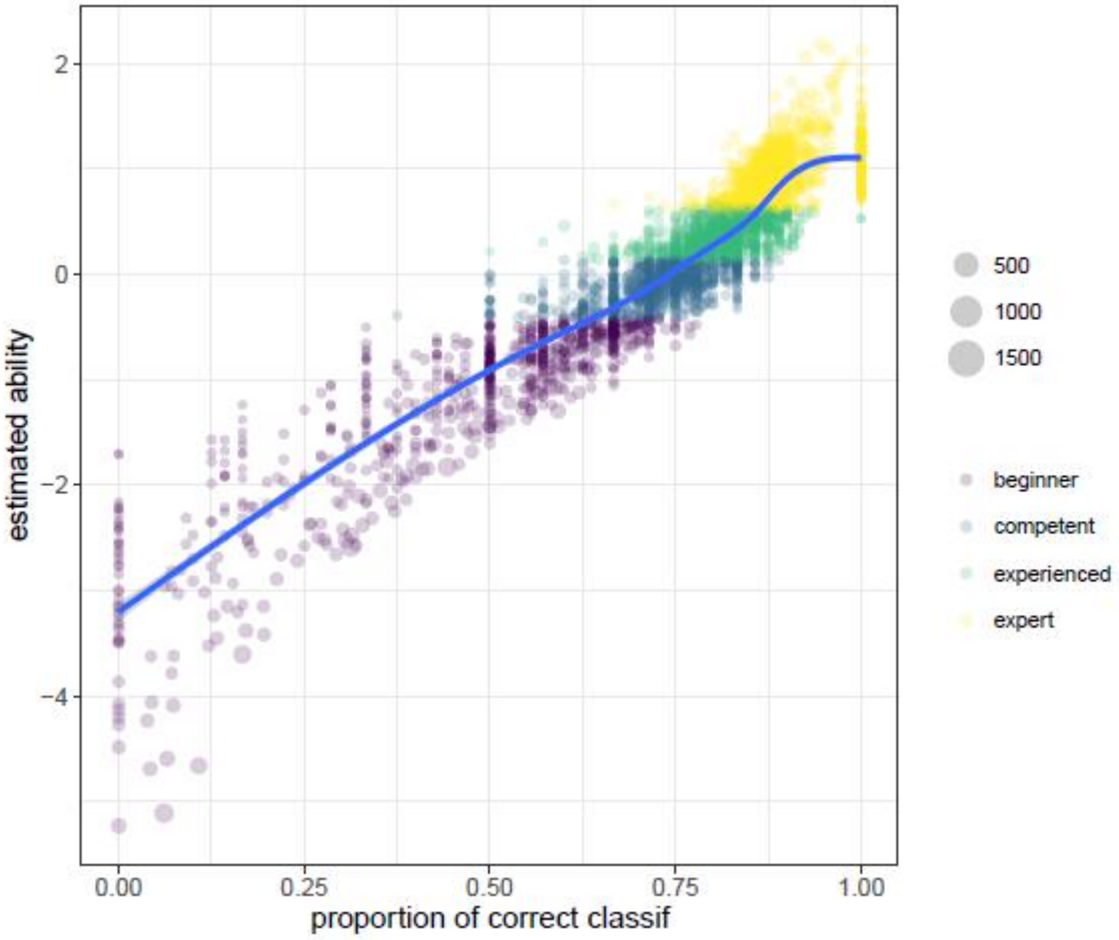
- Project “Snapshot-Serengeti” on *Zoouniverse :*
  Crowdsourcing to identify species in camera trap photos
- >1M images, 225 locations, 10.8M classifications from 28,000 users
- Gold standard dataset : 4,140 images classified by experts
- 50 species identified.



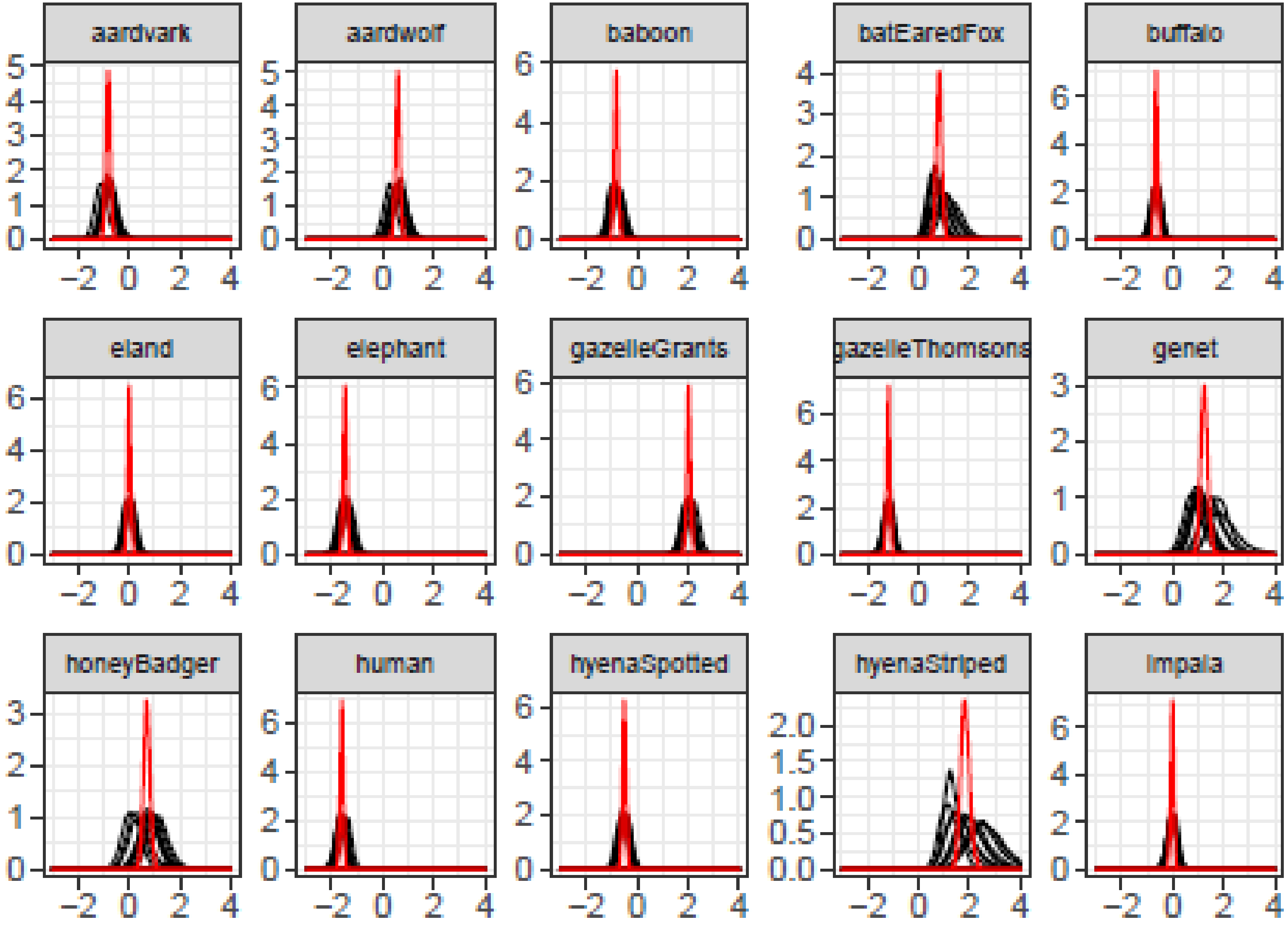
Big data computation

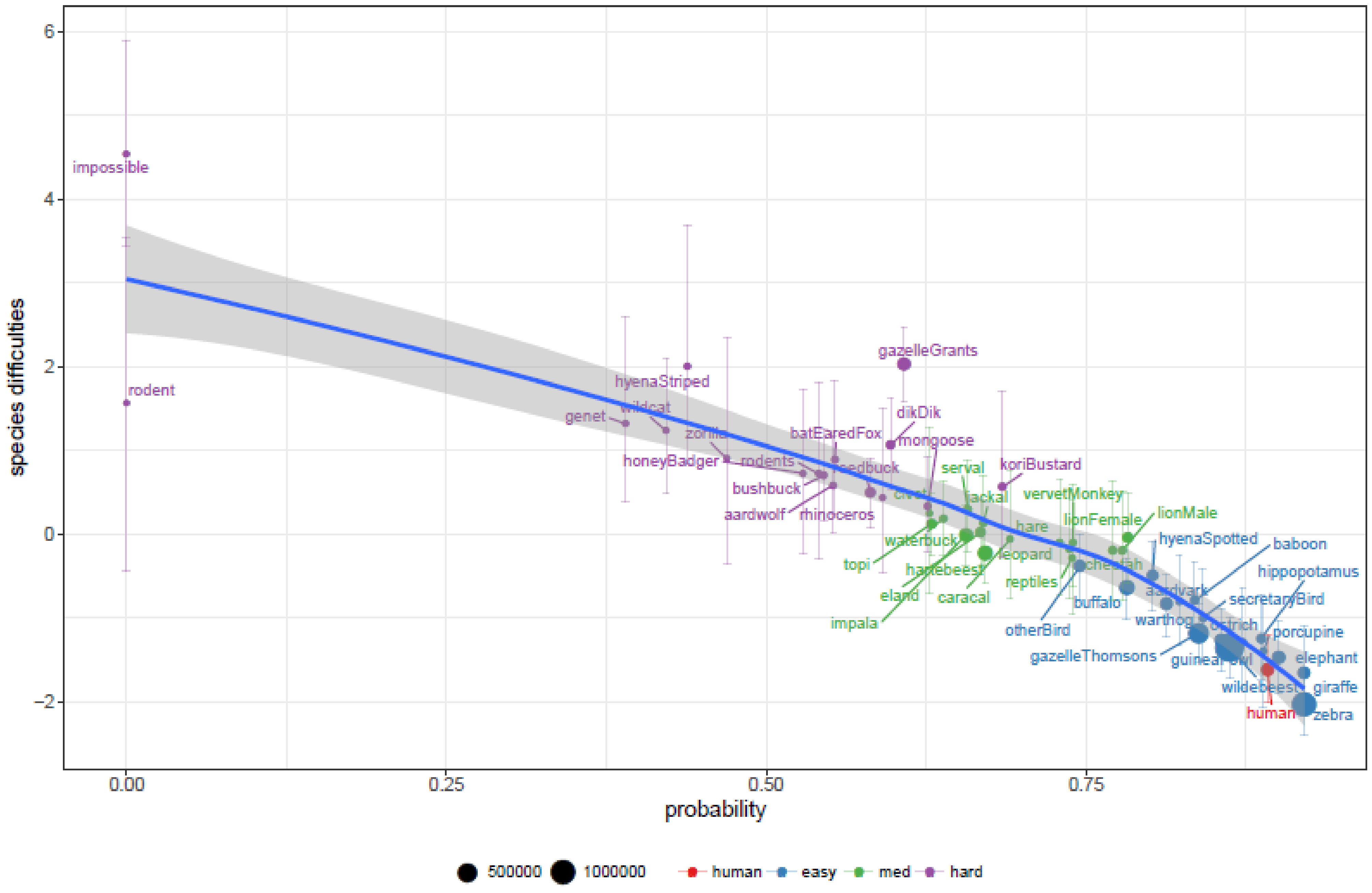
- How to handle datasets too large to be fit directly in one machine or even on a HPC?
- Divide-and-conquer or a divide-and-recombine approach:
  - split into multiple shards or subsets
  - fit models to independent subsets on independent machines
  - Combine subposterior estimates into global estimates using consensus Monte Carlo
  - Weighted averages of the posterior MCMC chains, weights prop. to 1/variance
- Alternative:
  - Divide users into 10 equal groups w.r.t. no. classifications
  - 10 shards with ~ 0.5M classifications per shard.
  - Fit shards in parallel with no communication.
  - Combine using stratified sampling principles.

Results – user abilities



Results - Posterior estimates of species difficulties




species difficulties
6
4
2
0
-2
probability
0.00
0.25
0.50
0.75
impossible
rodent
genet
wildcat
hyenaStriped
zorilla
honeyBadger
rodents
batEaredFox
reedbuck
bushbuck
aardwolf
rhinoceros
civet
gazelleGrants
dikDik
mongoose
koriBustard
serval
jackal
vervetMonkey
lionMale
lionFemale
hare
waterbuck
topi
hartebeest
eland
impala
caracal
leopard
reptiles
cheetah
buffalo
otherBird
hyenaSpotted
baboon
hippopotamus
aardvark
secretaryBird
warthog
ostrich
porcupine
gazelleThomsons
guineaFowl
elephant
wildebeest
giraffe
human
zebra
500000
1000000
human
easy
med
hard

Summary

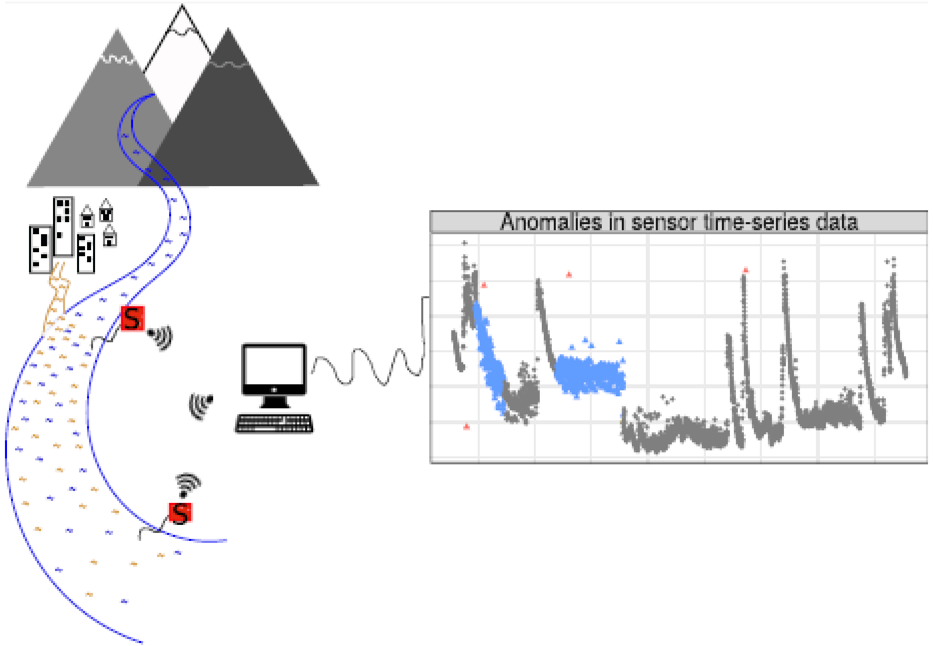
- New measurement error and item-response models for combining and adjusting crowdsourced geospatial data.

- Models outperform current popular approaches.

- Methods help to improve quality, accuracy and trust for crowdsourced data.

- Methods are scalable.

Case Study 2: Anomaly Detection in High-Dimensional Time Series



In-situ sensors **produce high-volume, high velocity data** describing fine-scale patterns, trends and extremes in space and near-real time

Data are prone to errors due to miscalibration, biofouling, and battery or other technical errors

- Manual QA/QC is too inefficient

Partner Organisations

- QLD Department of Environment and Science
- Southeast QLD Healthy Land & Water

Research Providers

- Universities: QUT, Monash, RMIT, University of Pau, University of Moratuwa, University of Alaska, EP Consulting

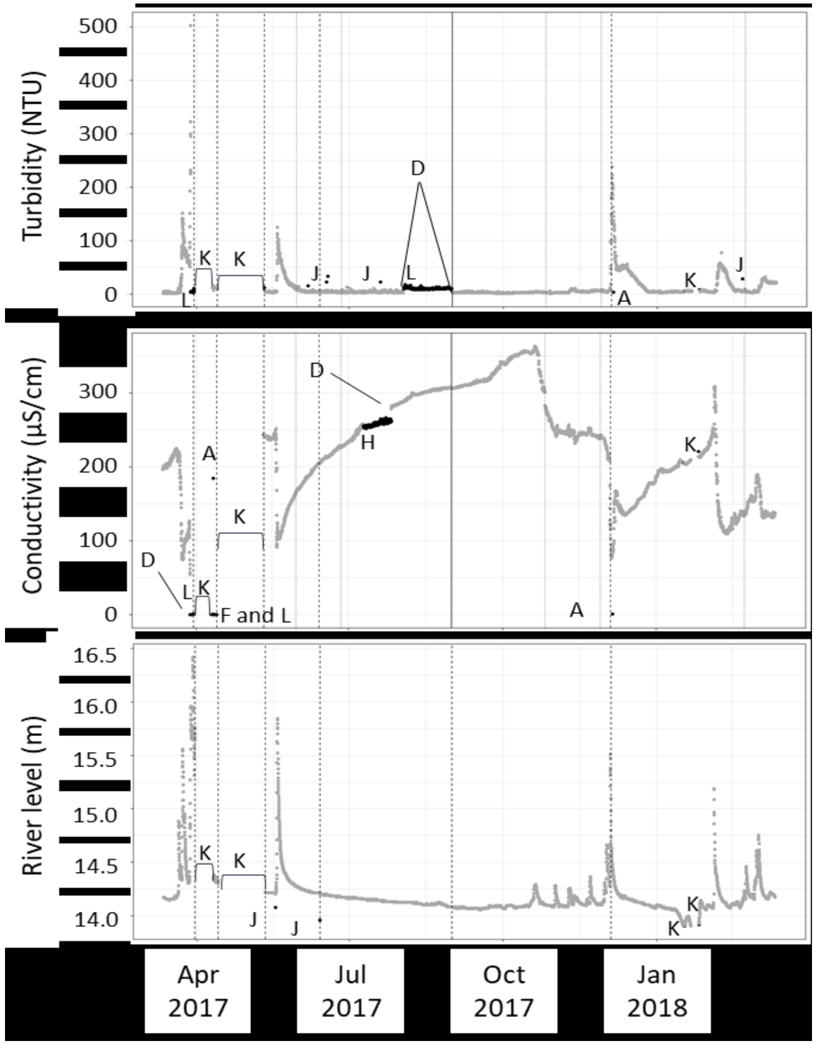
Other Collaborators:

- US National Ecological Observatory Network (NEON)
- US National Oceanographic & Atmospheric Administration (NOAA)

Types of anomalies

Type	Class
Large sudden spike	A
Low variability / persistent values	B
Constant offset (e.g. calibration error)	C
Sudden shifts	D
High variability	E
Impossible values	F
Out-of-sensor-range values	G
Drift	H
Clusters of spikes	I
Small sudden spike	J
Missing values	K

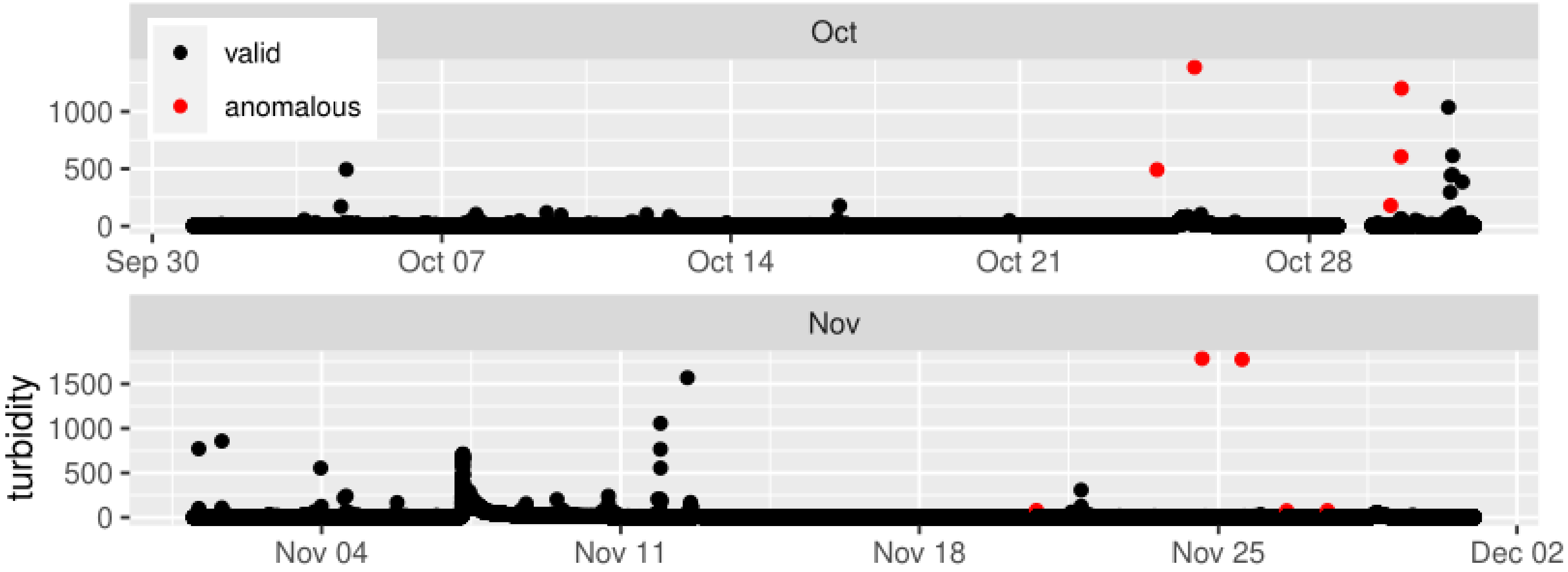
PIONEER RIVER



Cross-correlation approach

Can data from nearby sensors be used to detect anomalies?

- Turbidity ~ level, conductivity, and temperature from pairs of NEON sensors
- Estimate cross-correlation between up & downstream data, accounting for time lags
- Use GAM models to predict at downstream sensor



Framework for anomaly detection

Methods

- Rule-based:
  - e.g. ‘no negative values’
- Feature-based:
  - Consider patterns in multiple series (e.g., turbidity, conductivity)
  - e.g. HDOutliers, aggregated k-nearest neighbour (kNN-agg)
- Regression-based:
  - Fit model, e.g. ARIMA: $x_t = \beta' Z_t + \eta_t$; $\eta_t = ARIMA(p, d, q)$
  - Classify as ‘anomaly’ if the one-step-ahead prediction does not fall in the predictive interval.

Results

- Combination of methods facilitated the correct classification of impossible values, sudden isolated spikes and level shifts.
- Drift and periods of high variability still tended to be associated with high rates of false positives.

Artificial Neural Network (ANN) approach

- Two approaches:
  - Semi-supervised – train using only non-anomalous data; fit models with prediction errors; predict anomalous events as those observed to fall outside prediction intervals.
  - Supervised – train using labelled anomalous and non-anomalous data; generate probabilities that are binary-classified according to a predefined threshold.
- Used long short-term memory (LSTM) networks:
  - Type of Recurrent Neural Network (RNN), which is a special case of auto-regressive integrated moving average (ARIMA) and nonlinear auto-regressive moving average models
  - Structured in network units (memory blocks) composed of self-connected memory cells and three multiplicative units (“input”, “output”, “forget gates”)
- Explored different learning methods and hyperparameter values.
  - Used ‘keras’ software interfaced with R.

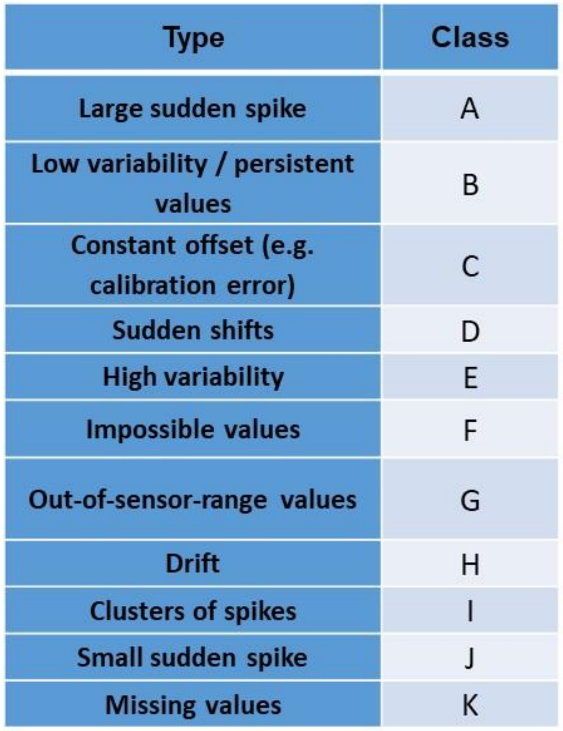
Artificial Neural Network (ANN) approach

- Calibrated the models using a Bayesian multiobjective optimization procedure:
  - Objective function based on combinations of accuracy, sensitivity, specificity, positive predictive value, negative predictive value
  - Iteratively improve the posterior distribution of functions (assuming a Gaussian process) associated with the variability of each hyperparameter, to maximise objective function.
  - Used ‘mlrMBO’ toolbox in R
- Selected and evaluated the “best” model for each water-quality variable, environment, and anomaly type.
- Identified influence of WQ variables on model performance using random forest (VI).

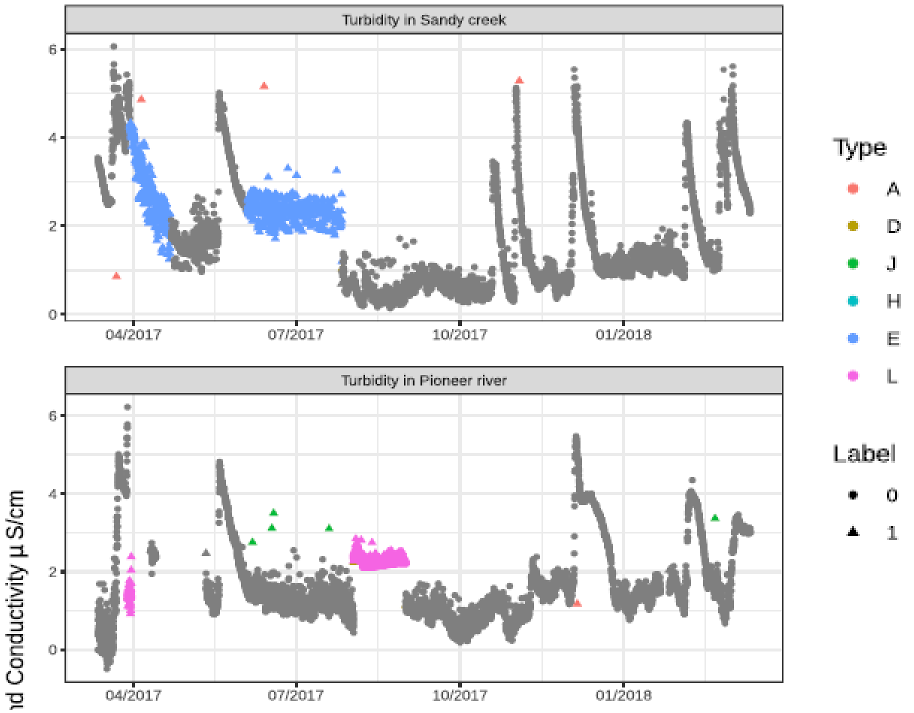
Artificial Neural Network (ANN) approach

Results

- Semi-supervised classification was better able to detect sudden spikes, sudden shifts, and small sudden spikes.
- Supervised classification had higher accuracy for predicting long-term anomalies associated with drifts and periods of otherwise unexplained high variability.

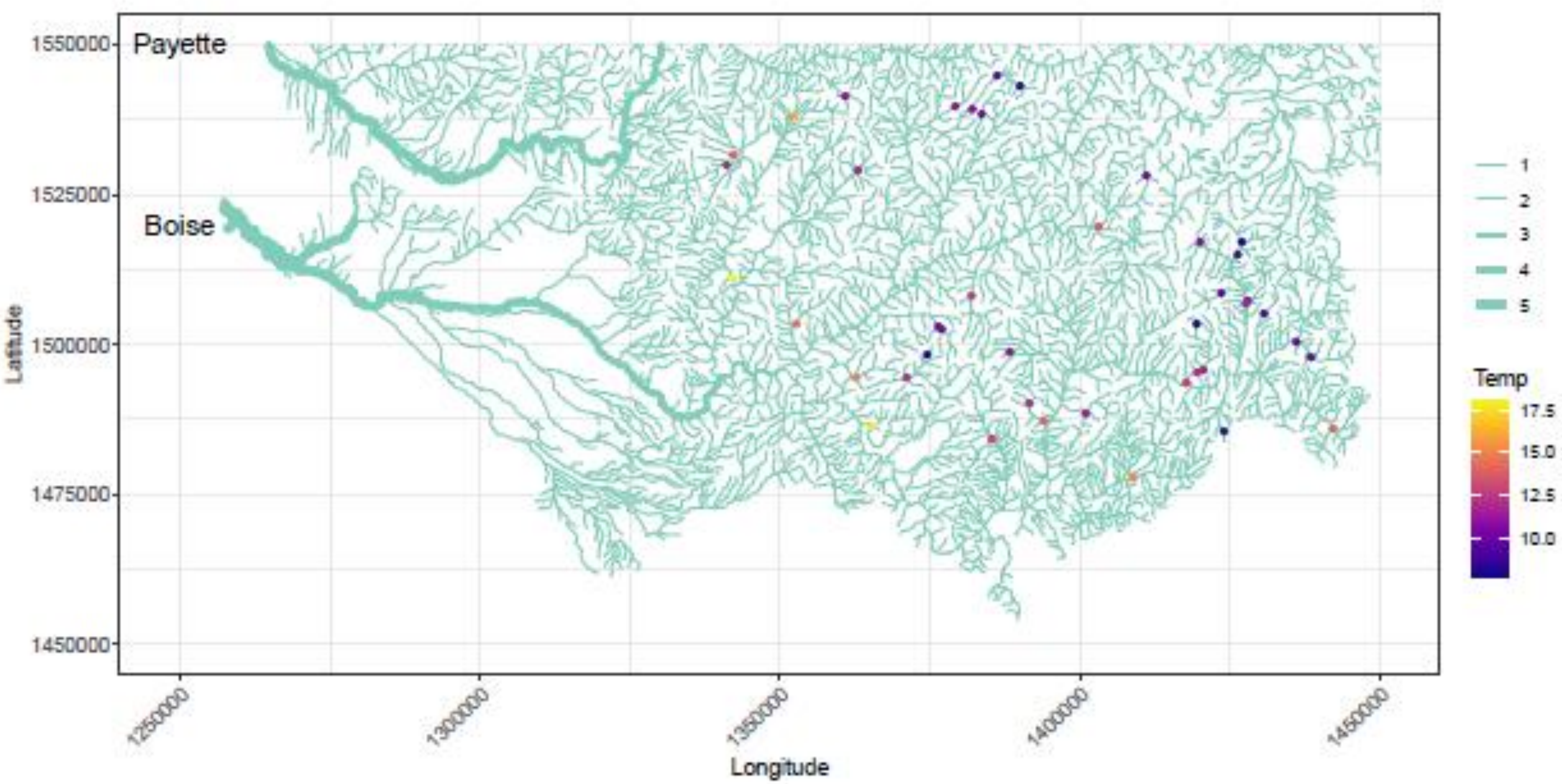


Type	Class
Large sudden spike	A
Low variability / persistent values	B
Constant offset (e.g. calibration error)	C
Sudden shifts	D
High variability	E
Impossible values	F
Out-of-sensor-range values	G
Drift	H
Clusters of spikes	I
Small sudden spike	J
Missing values	K



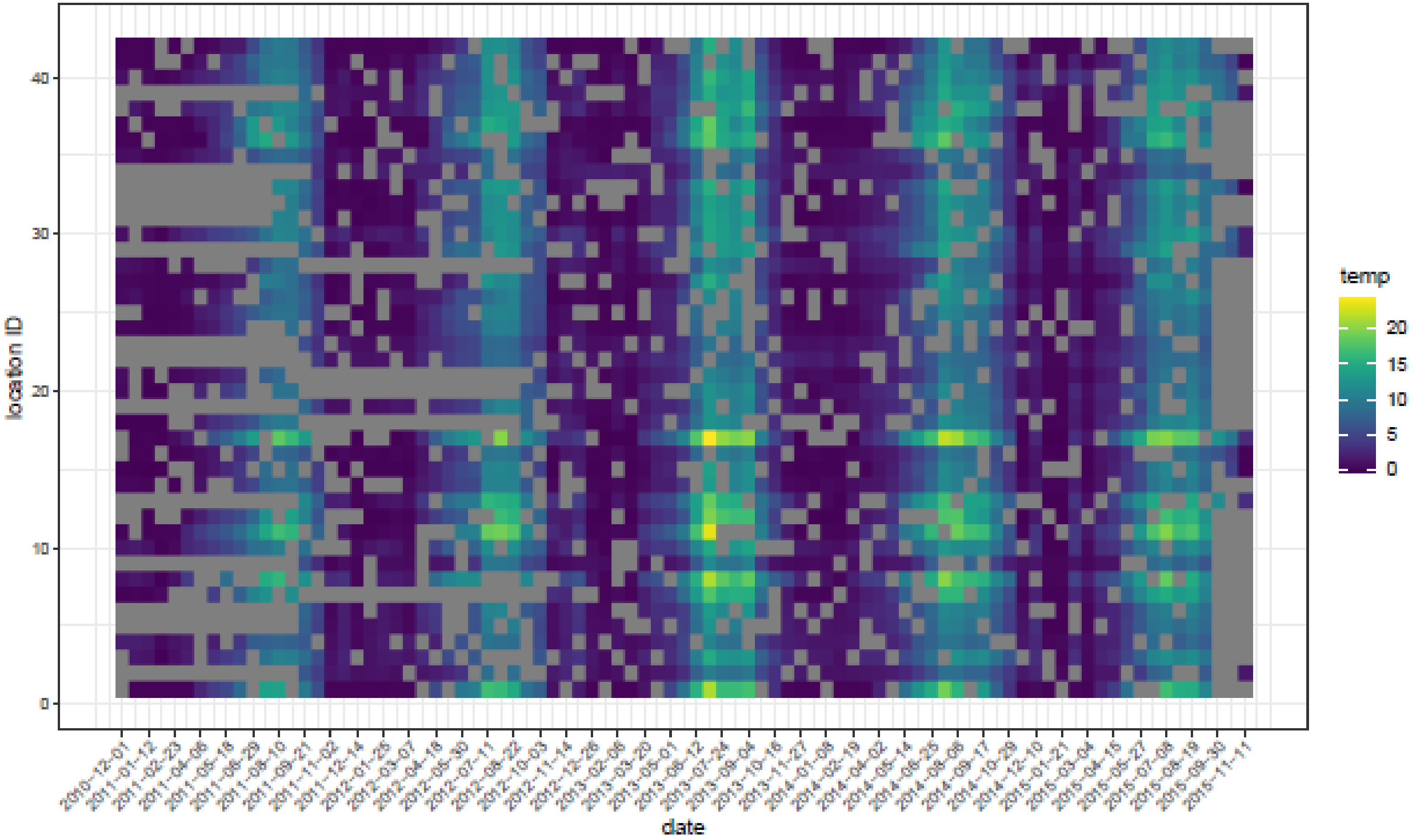
From Anomaly Detection to Spatio-Temporal model

- Temperature data for a stream network in the USA (Boise River)
- 42 data points – want to predict at every 1km, i.e. 1622 locations

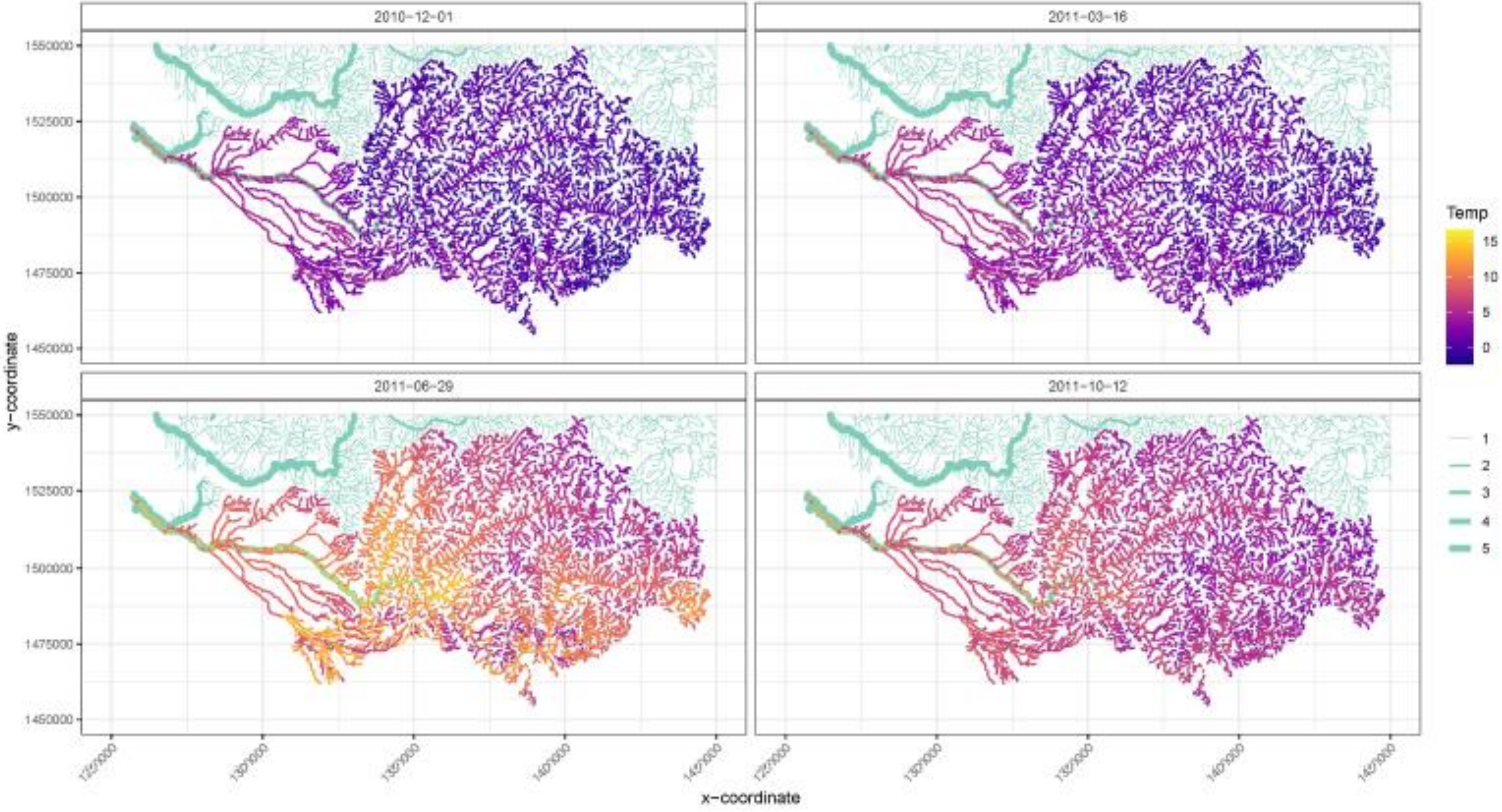


From Anomaly Detection to Spatio-Temporal model

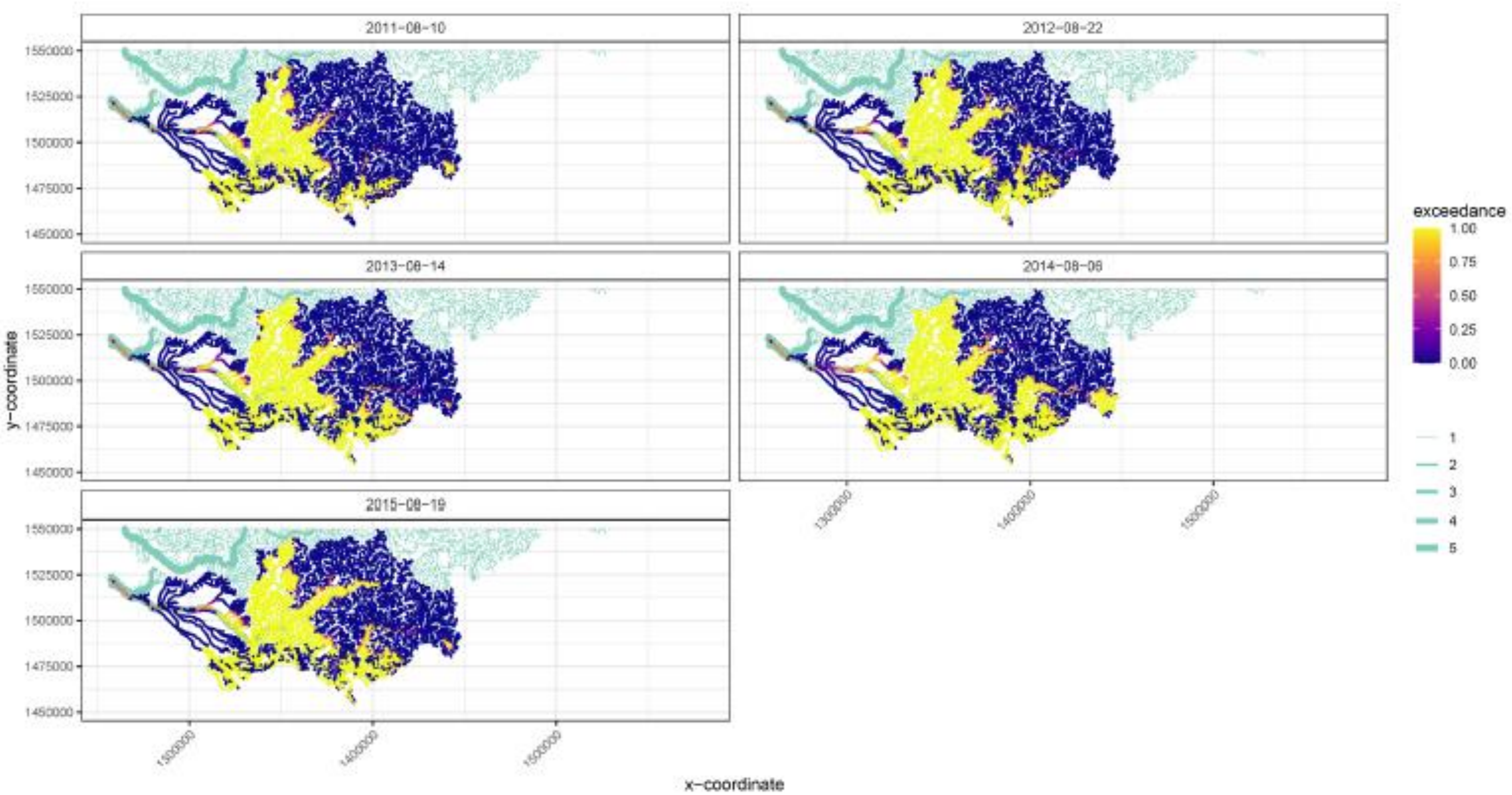
- Response: Subsampled water temperature data every 3 weeks over 5 years (87 time points)
- Covariates: air temperature, stream slope, elevation, cumulative drainage area



Spatio-temporal predictions of water temperature



Pr(mean water temperature > 13^{o})

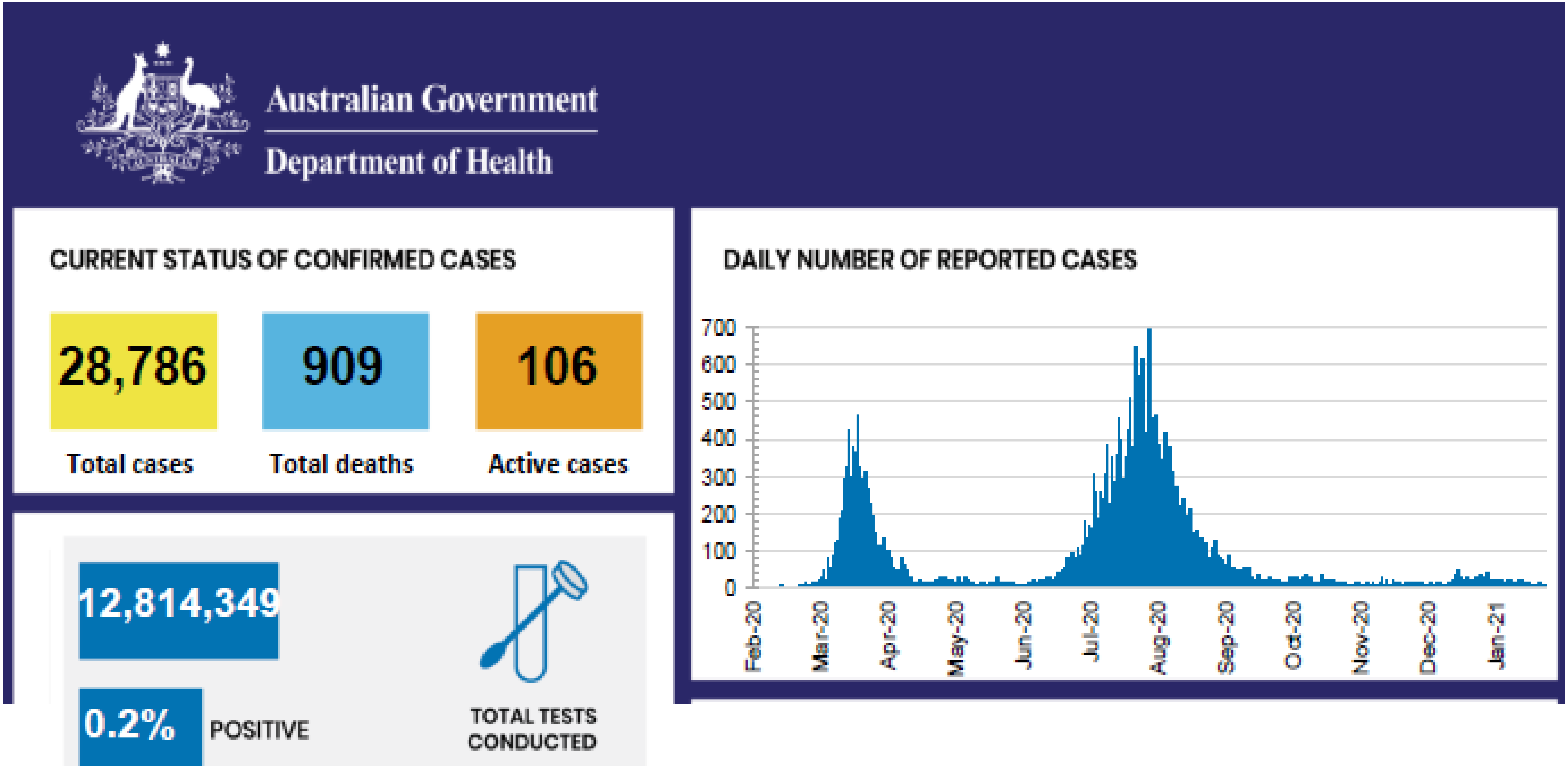


Case Study 2: Summary

1. Sensor data are pervasive.
2. Low-cost sensors have great potential for ecological monitoring, but they are prone to technical anomalies.
3. Statistical methods for anomaly detection can help to improve trust in the data.
4. Trustworthy data can provide strong data-focused insights.

Nothing remains the same from one moment to the next, you can't step into the same river twice.
Life – evolution – the whole universe of space/time, matter/energy – existence itself
is essentially change." - Ursula K. Le Guin

Case Study 3: Explainable models



Australian Government

Department of Health

CURRENT STATUS OF CONFIRMED CASES

28,786	909	106
Total cases	Total deaths	Active cases

12,814,349

0.2% POSITIVE

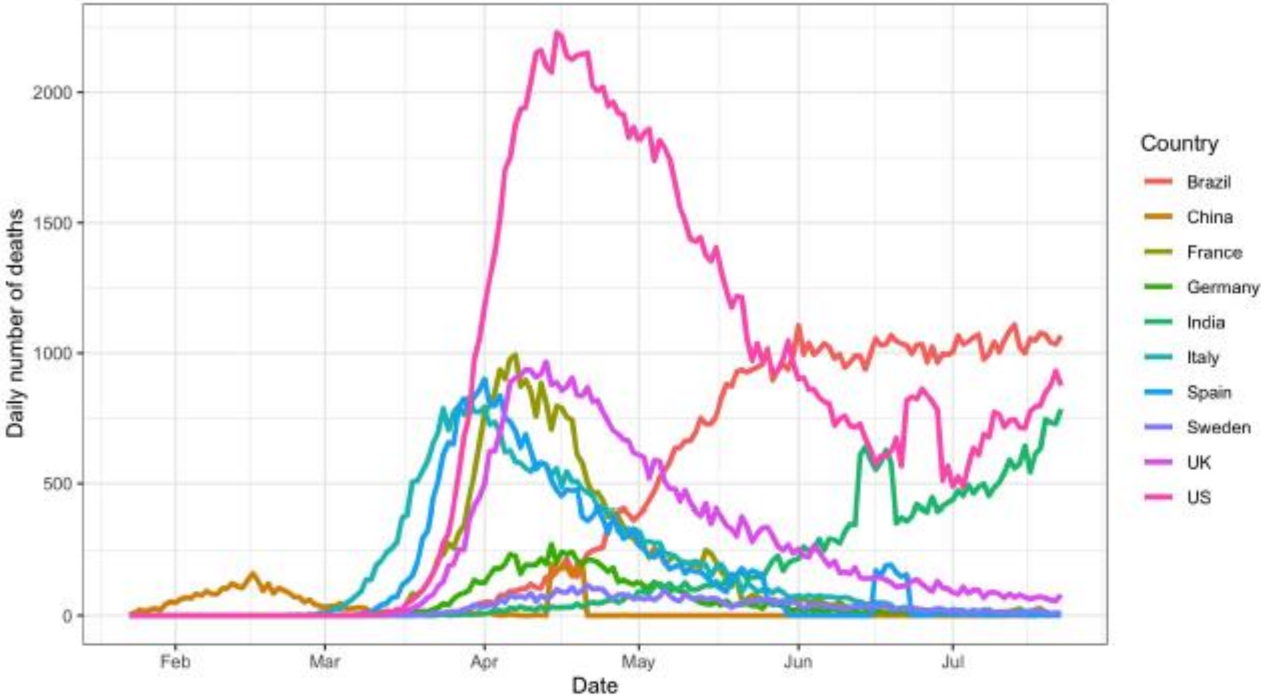
TOTAL TESTS CONDUCTED

DAILY NUMBER OF REPORTED CASES

Simple discrete-time self-exciting models can describe complex dynamics processes

- Hawkes process: past events influence the short-term probability of future events occurring.

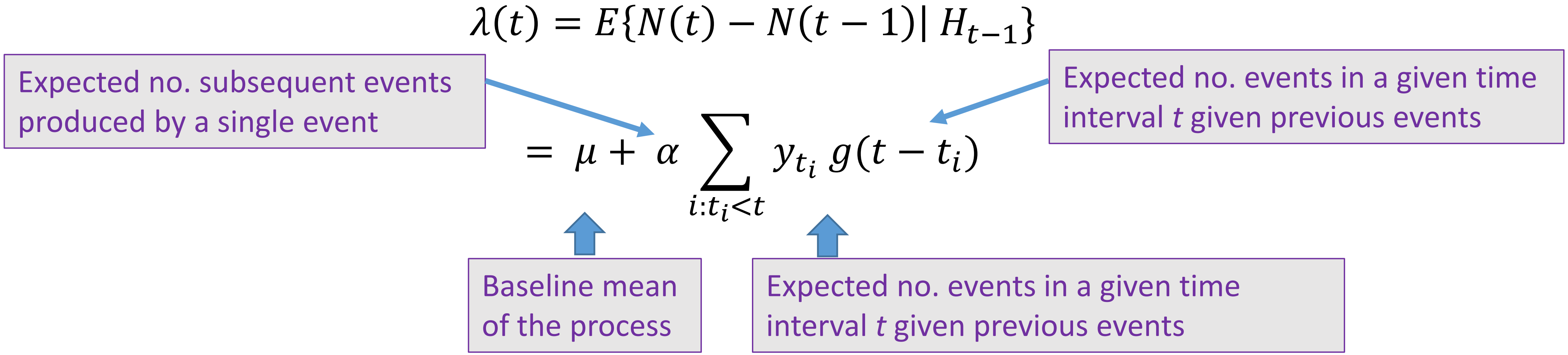
  - Neuroscience
  - Crime and terrorism
  - Seismic activity
  - Social media
  - Infectious diseases



Browning *et al.*, 2021
Warne *et al.*, 2020

Discrete-time Hawkes process

- Data (counts) y_t: number of events in a given time interval *s*: $y_t \sim \text{Poisson}(\lambda(t))$
- Conditional intensity function $\lambda(t)$: expected number of events that occur at time interval *t*, conditionally on the past.
- Number of events up to time *t*: N_t
- History of events up to but not including time *t*: H_{t-1}



$$\lambda(t) = E\{N(t) - N(t-1) | \, H_{t-1}\}$$

$$= \mu + \alpha \sum_{i:t_i<t} y_{t_i} \, g(t - t_i)$$

Expected no. subsequent events produced by a single event

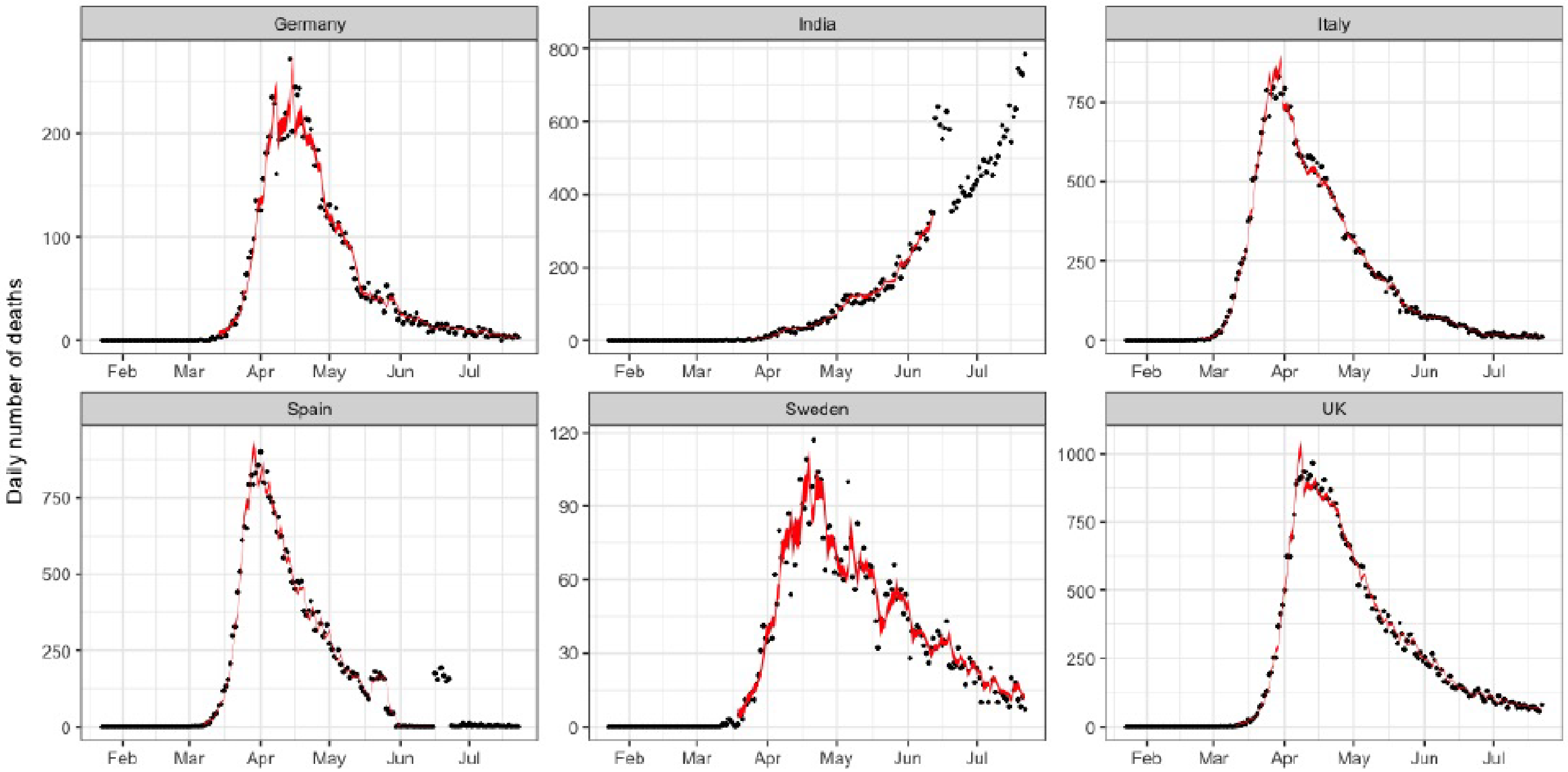
Expected no. events in a given time interval *t* given previous events

Baseline mean of the process

Expected no. events in a given time interval *t* given previous events

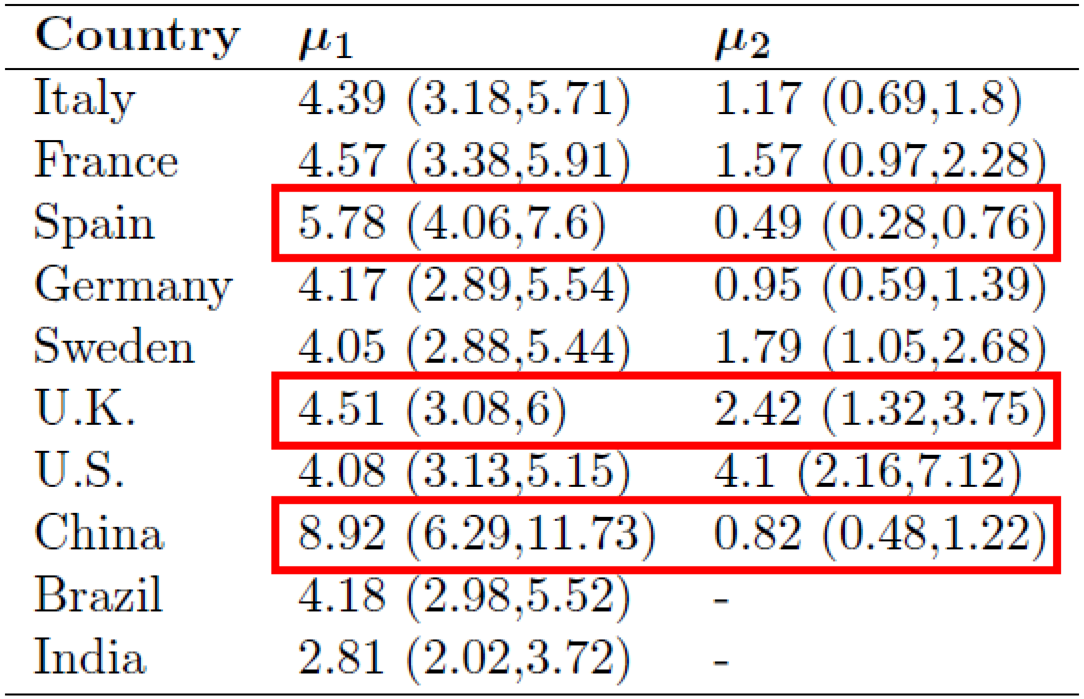
Results

1. Bayesian formulation
2. Include a change-point to describe different dynamics of the epidemic.
3. Estimate $(\mu_1, \alpha_1, \beta_1)$ and $(\mu_2, \alpha_2, \beta_2)$; estimate/predict trajectory.



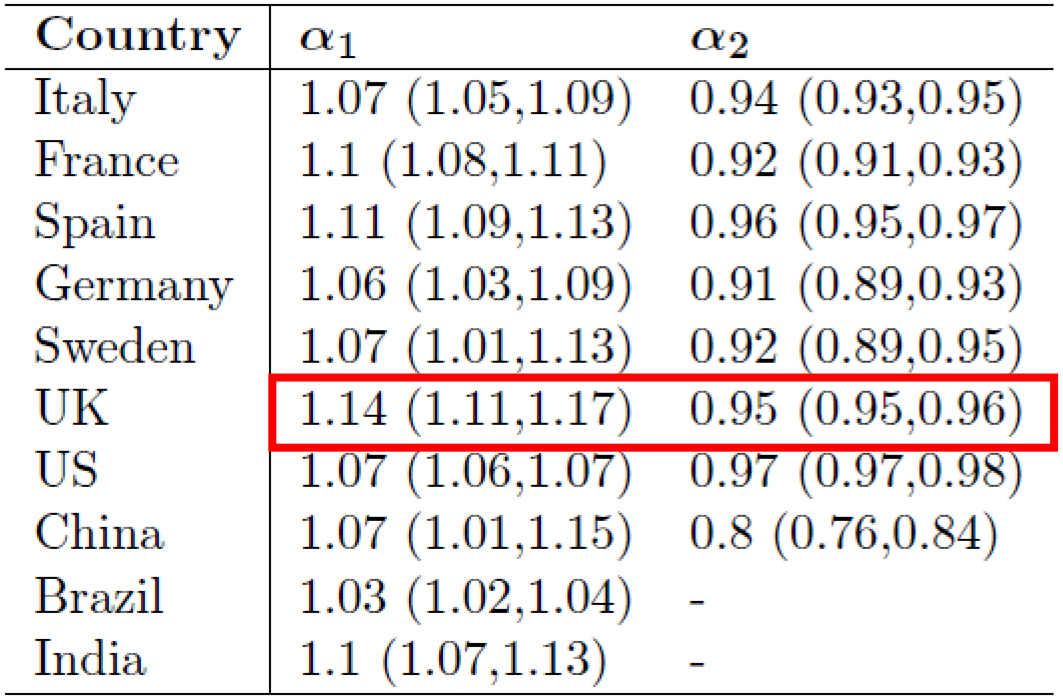
Results

Baseline mean of the process



Country	μ_1	μ_2
Italy	4.39 (3.18,5.71)	1.17 (0.69,1.8)
France	4.57 (3.38,5.91)	1.57 (0.97,2.28)
Spain	5.78 (4.06,7.6)	0.49 (0.28,0.76)
Germany	4.17 (2.89,5.54)	0.95 (0.59,1.39)
Sweden	4.05 (2.88,5.44)	1.79 (1.05,2.68)
U.K.	4.51 (3.08,6)	2.42 (1.32,3.75)
U.S.	4.08 (3.13,5.15)	4.1 (2.16,7.12)
China	8.92 (6.29,11.73)	0.82 (0.48,1.22)
Brazil	4.18 (2.98,5.52)	-
India	2.81 (2.02,3.72)	-

Expected number of events triggered by a single event



Country	α_1	α_2
Italy	1.07 (1.05,1.09)	0.94 (0.93,0.95)
France	1.1 (1.08,1.11)	0.92 (0.91,0.93)
Spain	1.11 (1.09,1.13)	0.96 (0.95,0.97)
Germany	1.06 (1.03,1.09)	0.91 (0.89,0.93)
Sweden	1.07 (1.01,1.13)	0.92 (0.89,0.95)
UK	1.14 (1.11,1.17)	0.95 (0.95,0.96)
US	1.07 (1.06,1.07)	0.97 (0.97,0.98)
China	1.07 (1.01,1.15)	0.8 (0.76,0.84)
Brazil	1.03 (1.02,1.04)	-
India	1.1 (1.07,1.13)	-

Summary

1. Sometimes, simple models can adequately capture complex dynamics.
2. The model can also quantify the dynamics of distinct phases in the pandemic.
3. We can also describe and compare country-specific dynamics.

Ongoing work:

Develop other models to analyse covid data.

Contribute further to developing methods to enhance trust in data and data science.

References

- Browning, R *et al.* (2021) Simple discrete-time self-exciting models can describe complex dynamic processes: a case study of COVID-19. Under review.
- Leigh C *et al.* (2019) A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. *Science of The Total Environment* 664, 885-898
- Leigh C *et al.* (2019) Using virtual reality and thermal imagery to improve statistical modelling of vulnerable and protected species. *PLoS ONE.*
- Santos-Fernandez E *et al.* (2020) Correcting misclassification errors in crowdsourced ecological data: A Bayesian perspective. arXiv.
- Santos-Fernandez E. *et al.* (2020) Bayesian item response models for citizen science ecological data. arXiv.
- Perez JR *et al.* (2020) Detecting technical anomalies in high-frequency water-quality data using Artificial Neural Networks. *Environmental Science & Technology*
- Virtual Reef Diver https://virtualreef.org.au