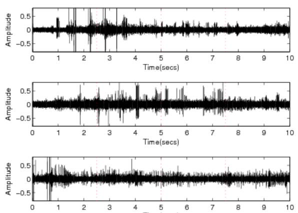
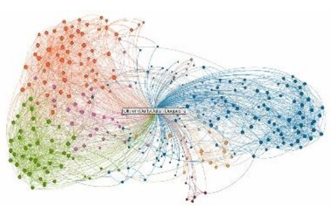
Bayesian learning for decision-making in the “big data” era

Kerrie Mengersen

Distinguished Professor of Statistics, QUT, Australia
D/Director, ARC Centre of Excellence in
Mathematical and Statistical Frontiers

“The Fourth Industrial Revolution”

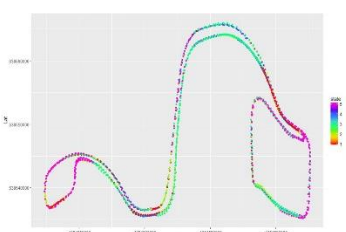


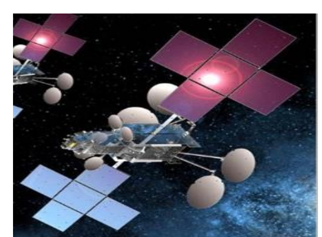


convergence of computing, data, artificial intelligence and universal connectivity

“raw data, no matter how extensive, are useless without a model.”
– Nate Silver

	age	sex
)	21.79603	female
)	21.71116	female
)	21.27584	male
)	27.37577	female
)	19.66598	male
)	24.37509	female

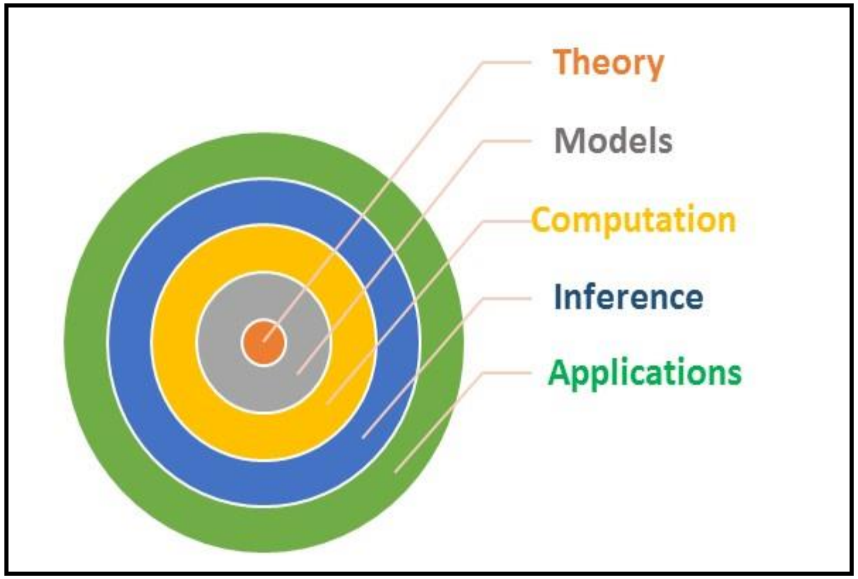




Bayesian Modelling

$$p(\theta|Y) = p(Y|\theta)\, p(\theta) \,/\, p(Y)$$

$$p(\theta|Y) = p(\theta)\, p(Y|\theta) \,/\, p(Y)$$



Meeting the challenge: “New Bayes”

Models:

- Probabilistic
- Regularised
- Flexible
- Robust
- Transferable
- Adaptive

Computation:

- Scalable (parallelisable)
- Subsampling
- Pre-computation
- Approximations

Inference:

- Estimation
- Optimisation
- Uncertainty quantification
- Testing
- Model averaging

"In the past ten years, it's hard to find anything that doesn't advocate a Bayesian approach." –Nate Silver

Focus in this presentation:

Methods:

- Data-focused modelling
- Making informed decisions
- Role and formulation of priors
- Visualisation
- Computation
- Diverse data sources

Applications:

- Air quality and health: personalised decision support
- Spatial patterns in cancer: an Australian Cancer Atlas
- Modelling complex queues
- Understanding gene expression
- Monitoring the health of the Great Barrier Reef

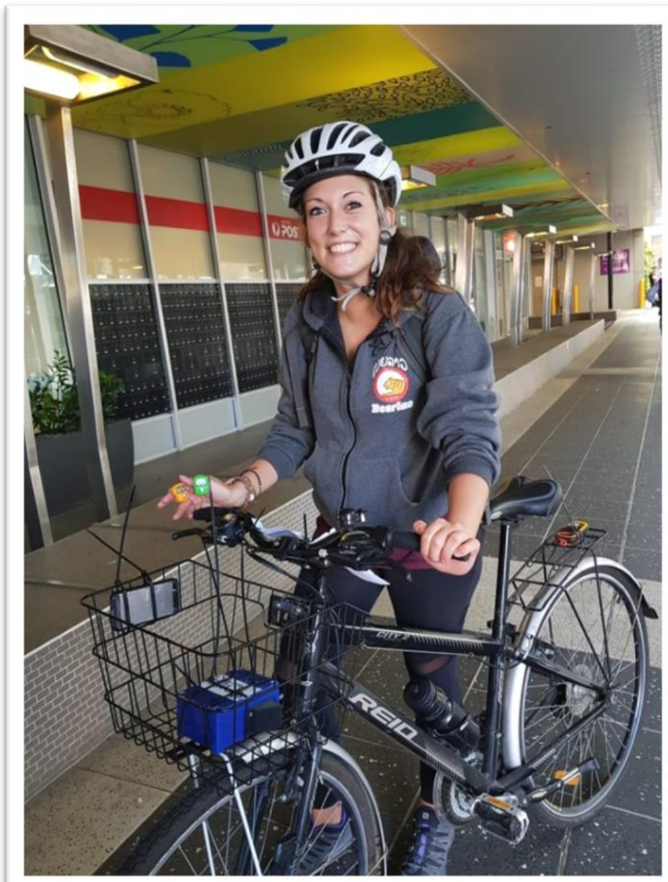
Air quality and health

Air quality and health

Personalised environmental health app: pilot study

L. Dawkins, D. Willimson, KM(2019)
Where is the clean air? A Bayesian decision framework for personalised cyclist route selection.

Current mobile apps and websites gives “one size fits all” guidance.



Demand is for personalised air quality guidance, e.g. optimal cycle route from A to B.

General approach

1. Develop a Bayesian spatio-temporal model to create a map of PM2.5 exposures along potential routes, based on high resolution data from mobile air quality sensors.

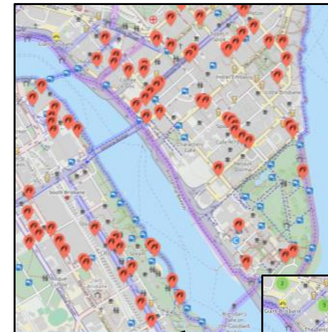
2. Use a Bayesian decision framework to create a user-specific multiattribute utility function

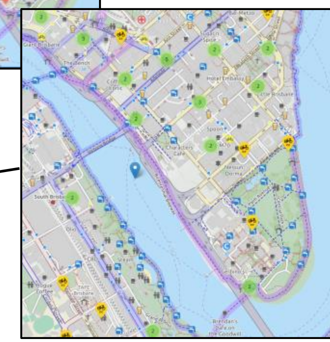
   Elicit a user's journey preferences, regarding health impact of exposure to PM2.5, journey time and journey enjoyment, via an R shiny web app.

3. Identify personalised optimal route as the one that maximises the expectation of this function.

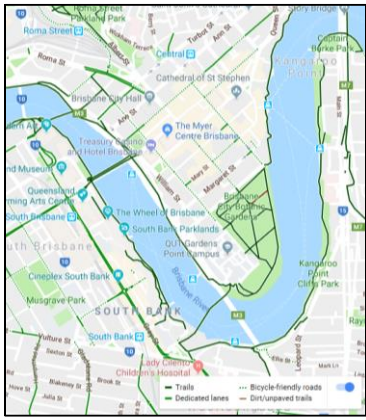
Data Collection: Where?

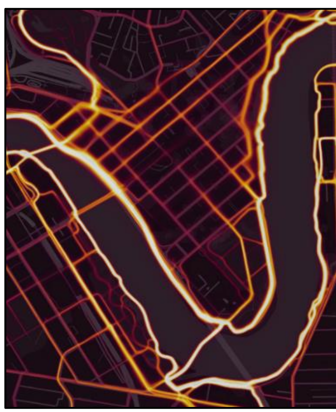
Encompass bike racks & CityCycle stations

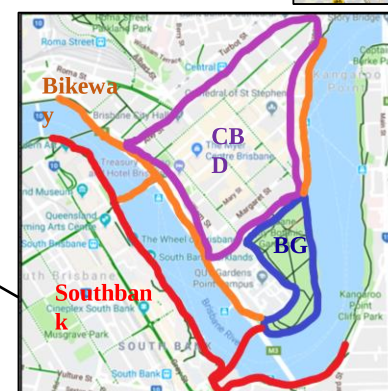




Cycle friendly routes





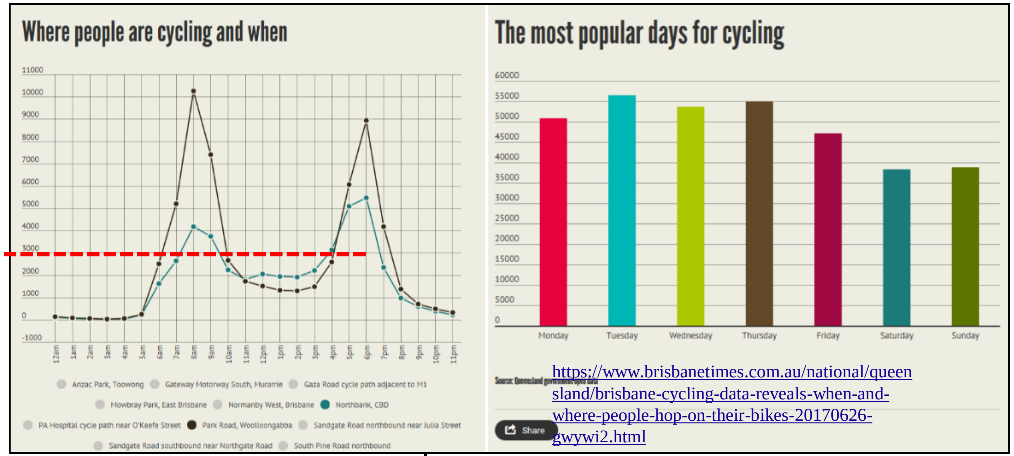


Different sections of CBD

“Hot” on STRAVA heat map

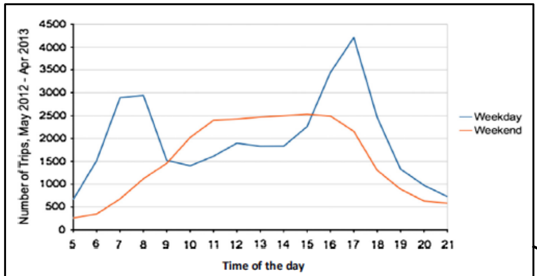
Data Collection: When?

Brisbane cycling count data



https://www.brisbanetimes.com.au/national/queensland/brisbane-cycling-data-reveals-when-and-where-people-hop-on-their-bikes-20170626-gwywi2.html

CityCycle paper: Mateo-Babiano et al. (2016)
How does our natural and built environment affect the use of bicycle sharing? Transportation Research Part A, 94:295–307



(b) Trip frequency

Frequency:

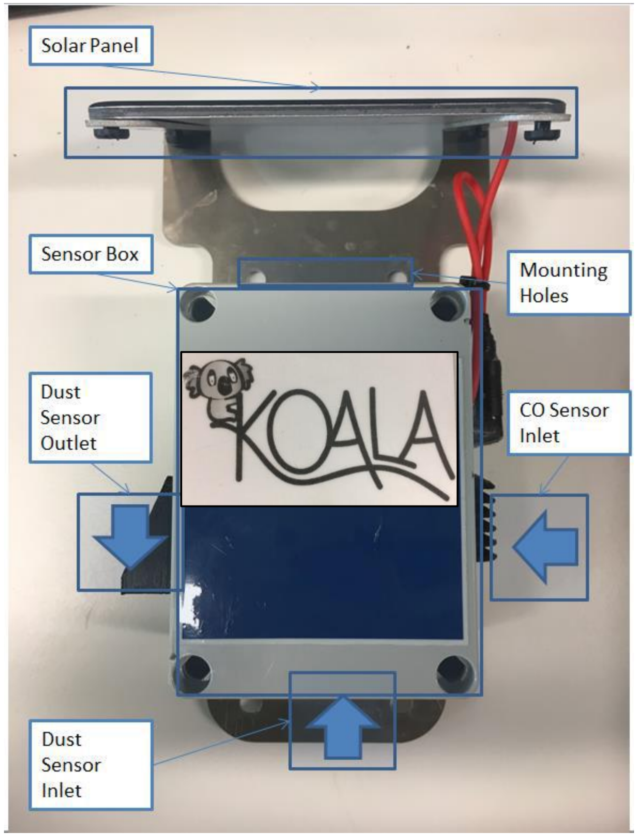
Literature on modelling mobile sensor data (e.g. Van den Bossche, J et al. (2015). Mobile monitoring for mapping spatial variation in urban air quality: Development and validation of a methodology based on an extensive dataset. Atmospheric Environment, 105:148-161.

along two fixed routes (2 and 3 km long). Large gradients over short distances and differences up to a factor of 10 in mean BC concentrations aggregated at a resolution of 20 m are observed. Mapping at such a high resolution is possible, but a lot of repeated measurements are required. After computing a trimmed mean and applying background normalisation, depending on the location 24–94 repeated measurement runs (median of 41) are required to map the BC concentrations at a 50 m resolution with an uncertainty of 25%. When relaxing the uncertainty to 50%, these numbers reduce to 5–11 (median of 8) runs. We conclude that mobile monitoring is a suitable approach for mapping urban air quality at a high spatial resolution, and can provide insight into the spatial variability that would not be possible with

Date	Trip 1	Trip 2	Trip 3	Trip 4
Thurs 10th May	7:00-8:30 Pair: Laura & Steve	8:30-10:00 Pair: Leah & Matt	15:30-17:00 Pair: Jess C & Victor	17:00-18:30 Pair: Tarun & Laura
Sun 13th May	10:00-11:30 Pair: Laura & Phill	11:30-13:00 Pair: Lidwina & Tarun		
Tues 15th May	7:00-8:30 Pair: Laura & Steve	8:30-10:00 Pair: Jess C & Victor	15:30-17:00 Pair: Omar & Miles	17:00-18:30 Pair: Lidwina & Jessie
Thurs 17th May	7:00-8:30 Pair: Laura & Tarun	8:30-10:00 Pair: Victor & Laura	15:30-17:00 Pair: Jessie & Miles	17:00-18:30 Pair: Lidwina & Edgar
Sun 20th May	10:00-11:30 Pair: Laura & Phill	11:30-13:00 Pair: Kerrie & Gerard		
Tues 22nd May	7:00-8:30 Pair: Laura & Steve	8:30-10:00 Pair: Jess C & Victor	15:30-17:00 Pair: Leah & Matt	17:00-18:30 Pair: Laura & Jessie
Thurs 24th May	7:00-8:30 Pair: Laura & Natalia	8:30-10:00 Pair: Jessie & Lidwina	15:30-17:00 Pair: Edgar & Natalia	17:00-18:30 Pair: Leah & Matt

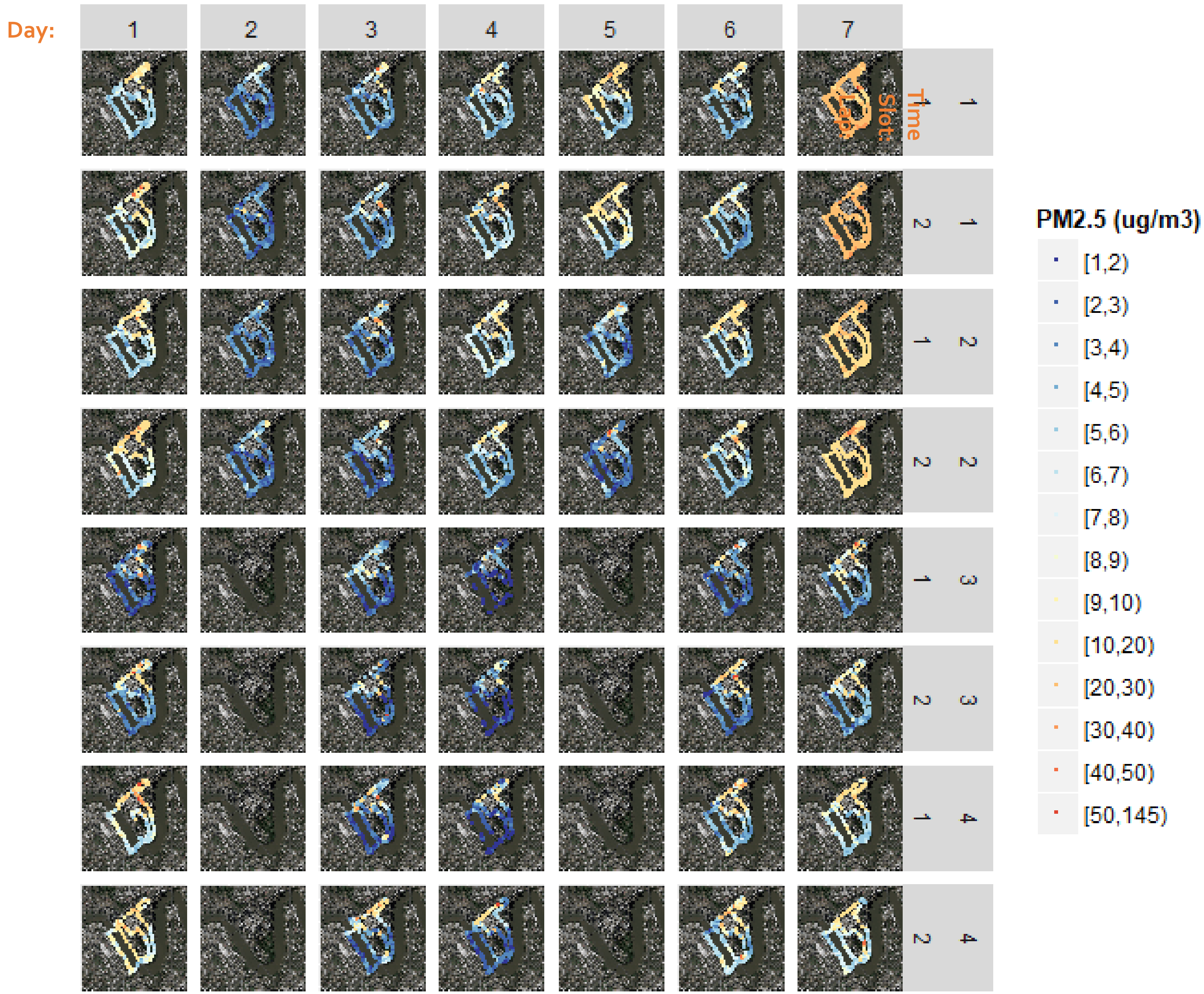
Total of 48 laps of the route over 24 time slots

The Air Quality Sensor: KOALA



- **Measures PM2.5 every 5 seconds** using the Dust Sensor
- **Calibration:** carried out tests in Botanic Gardens to ensure speed did not affect the PM2.5 measurement, found it was important to have the dust sensor inlet facing backwards

Data: All laps



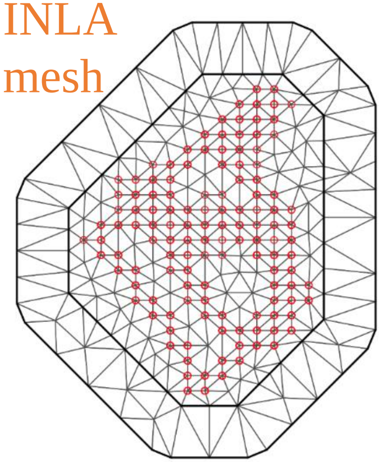
Developing the model

- Covariate selection: expert and cyclist insight

- Exploratory data analysis to identify most important
    - At time slot level
        - PM2.5 in South Brisbane preceeding hour
        - Temperature in CBD * Lane Type (on road/off road)
        - Wind direction in CBD (West/East) * Lane Type
        - Wind speed in South Brisbane * Lane Type
        - Humidity in South Brisbane
    - At observation level
        - Traffic counts at 4 locations (every 30 mins) * Time Of Day * Lane Type
        - Distance to major junction
        - Distance to river

Developing the model: grid cell resolution

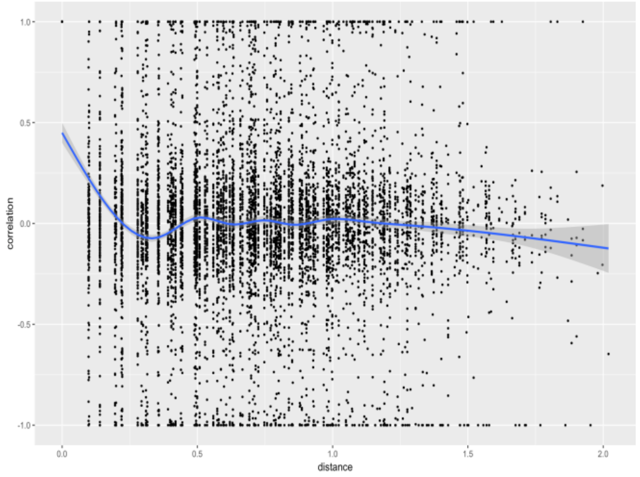
INLA mesh



For day d=1,…,D, time slot t=1,…,T_d, grid cell s=1,.

$$Z_{dts} = \beta_0 + \beta X_{dts} + \epsilon_{dts} + \eta$$

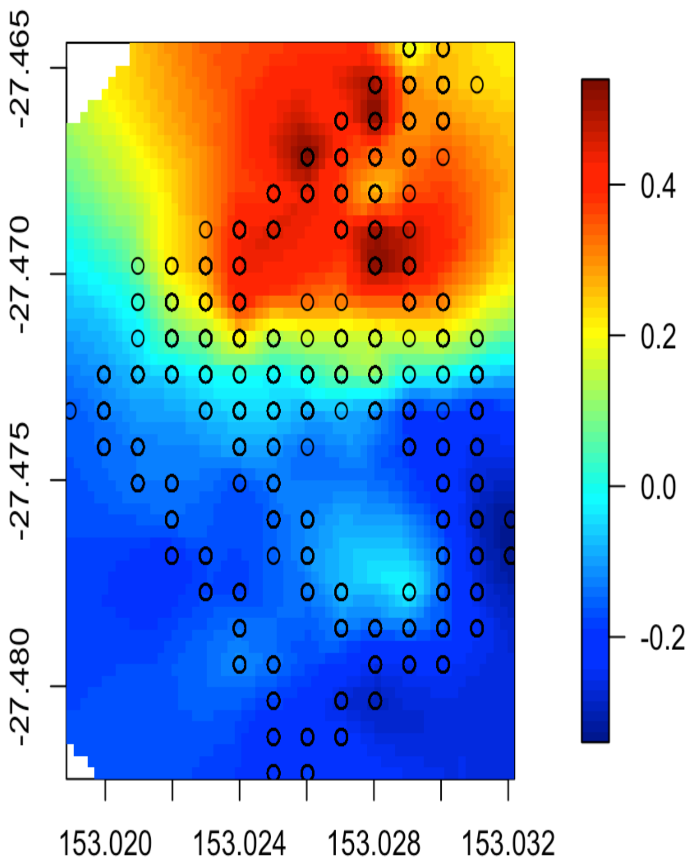
- Z_{dts} is the 100x100m grid cell averaged log(PM2.5+1) for day d, time slot t, grid cell s
- η is a spatial random effect with exponential spatial correlation function.



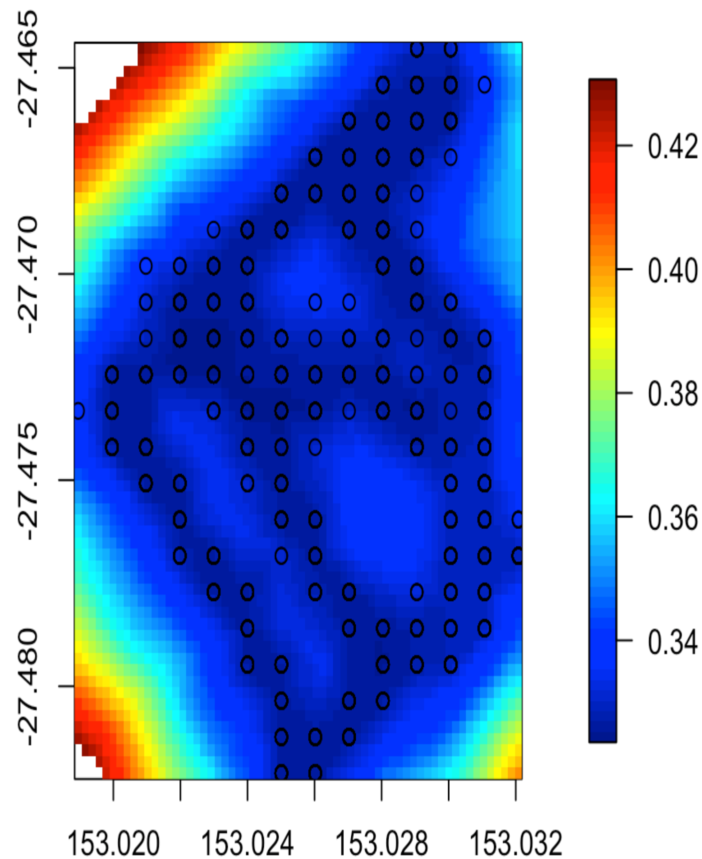
- Specify prior means for fixed effects as 1 or -1 for effects we have insight about the sign of the relationship, e.g. stronger winds => lower PM2.5
- Exponential spatial covariance structure

Results

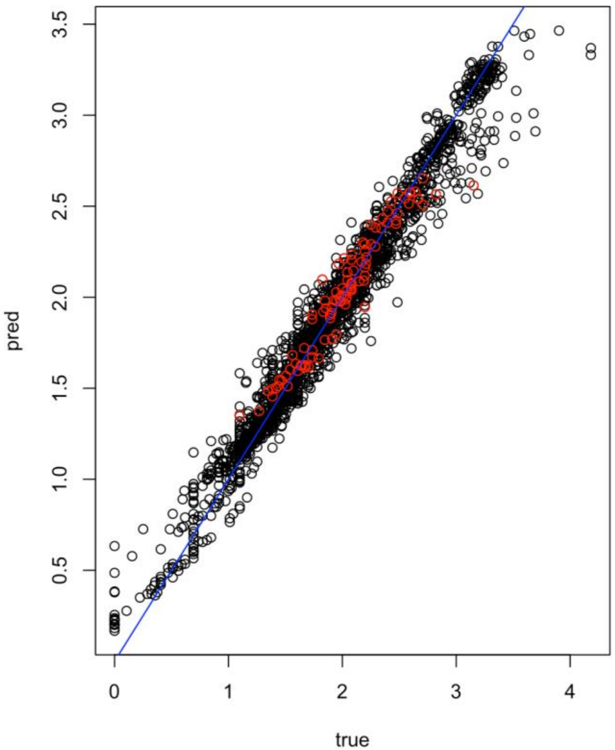
Mean of random field



SD of random field



Predictive capability

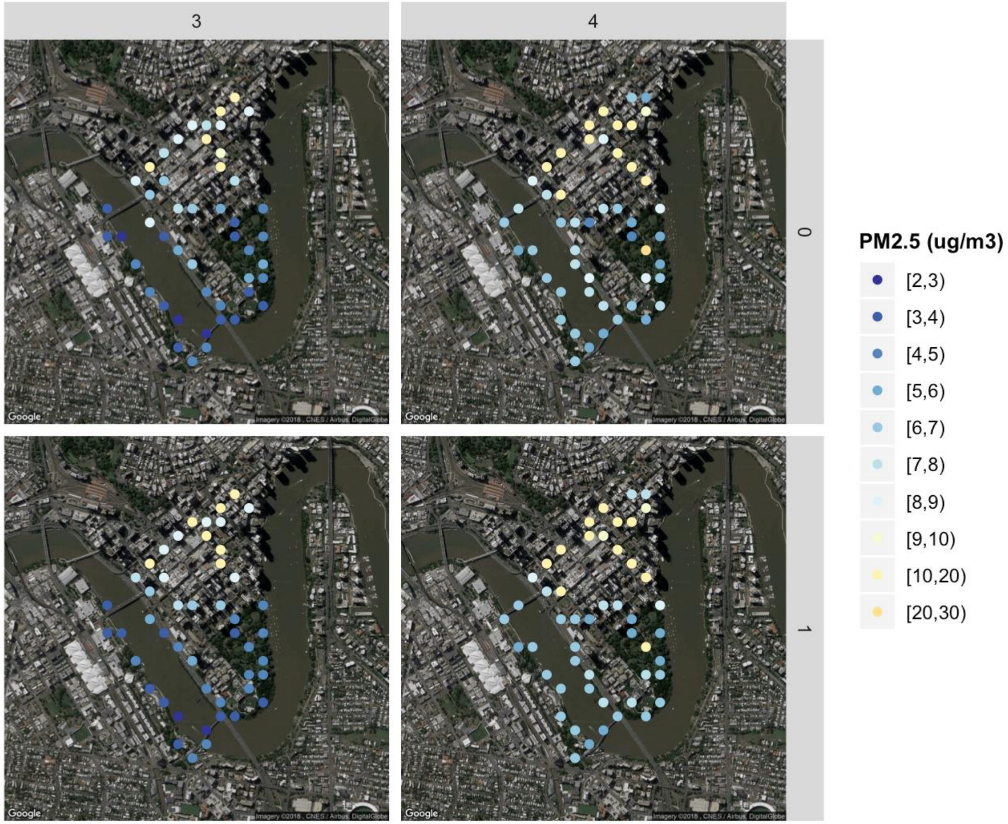


Red = data left out of model

Predicting data left out of model
(day 7 afternoon – timeslots 3 and 4)

True:

Predicted:

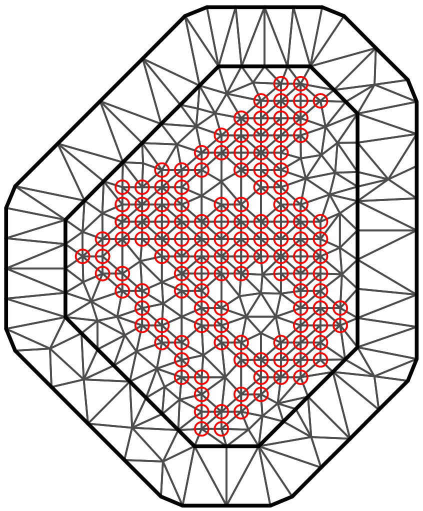


Developing the model: observation level resolution

For day d=1,…,D, time slot t=1,…,T_d, lap =1,2, grid cell s=1,…,S, observation in grid cell s, j=1,…J_{dtls}

X: covariates at the same resolution as the observations (traffic)

Z: covariates at the same resolution as the spatial-temporal grid



INLA mesh

$$Y_{dtsj} = \gamma X'_{dtsj} + Z_{dts} + \xi_{dtsj}$$

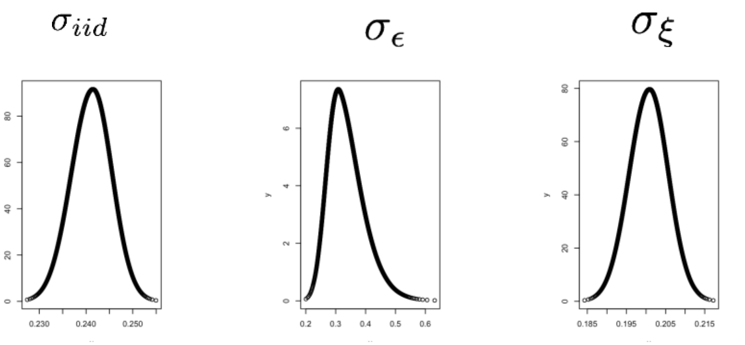
$$Z_{dts} = \beta_0 + \beta X_{dts} + \epsilon_{dts} + \eta$$

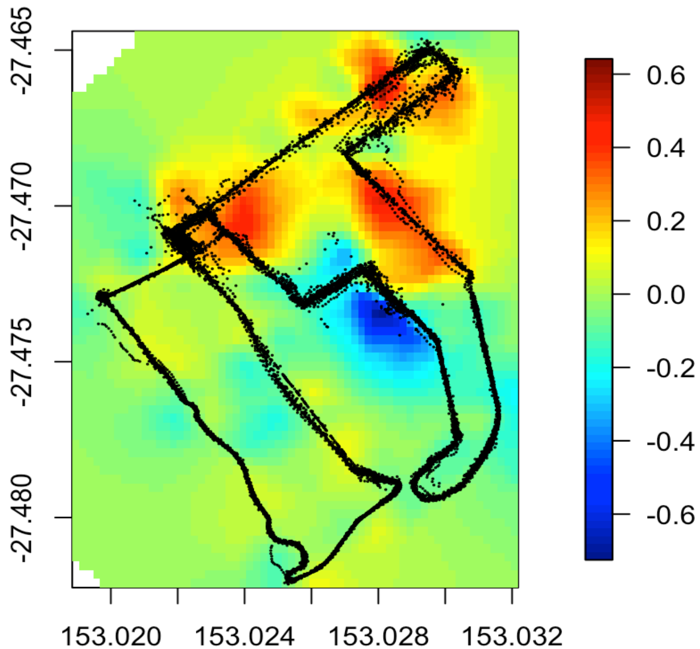
$$\xi_{dtsj} \sim N(0, \sigma_\xi^2) \qquad 1/\sigma_\xi^2 \sim logGamma(a, b)$$

$$\epsilon_{dts} \sim N(0, \sigma_\epsilon^2) \qquad 1/\sigma_\epsilon^2 \sim logGamma(a, b)$$

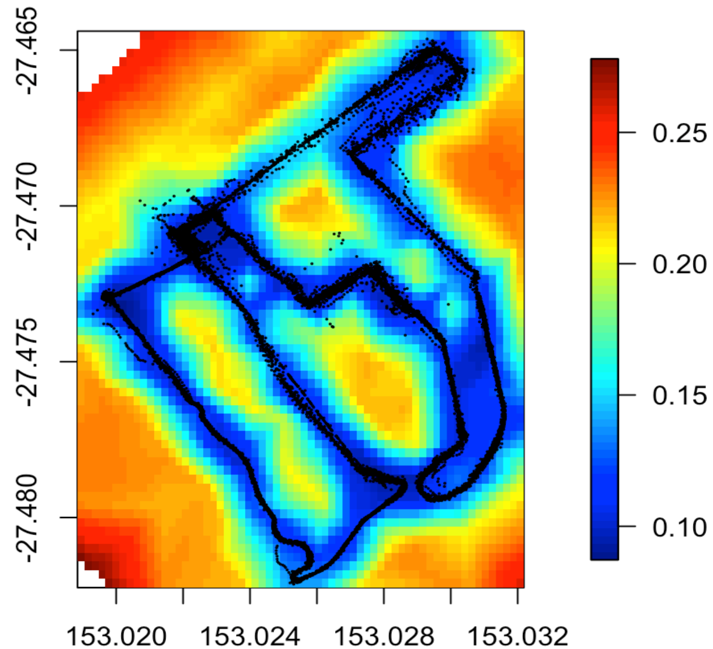
Same mesh, alter projection matrix to map to observed locations

Results





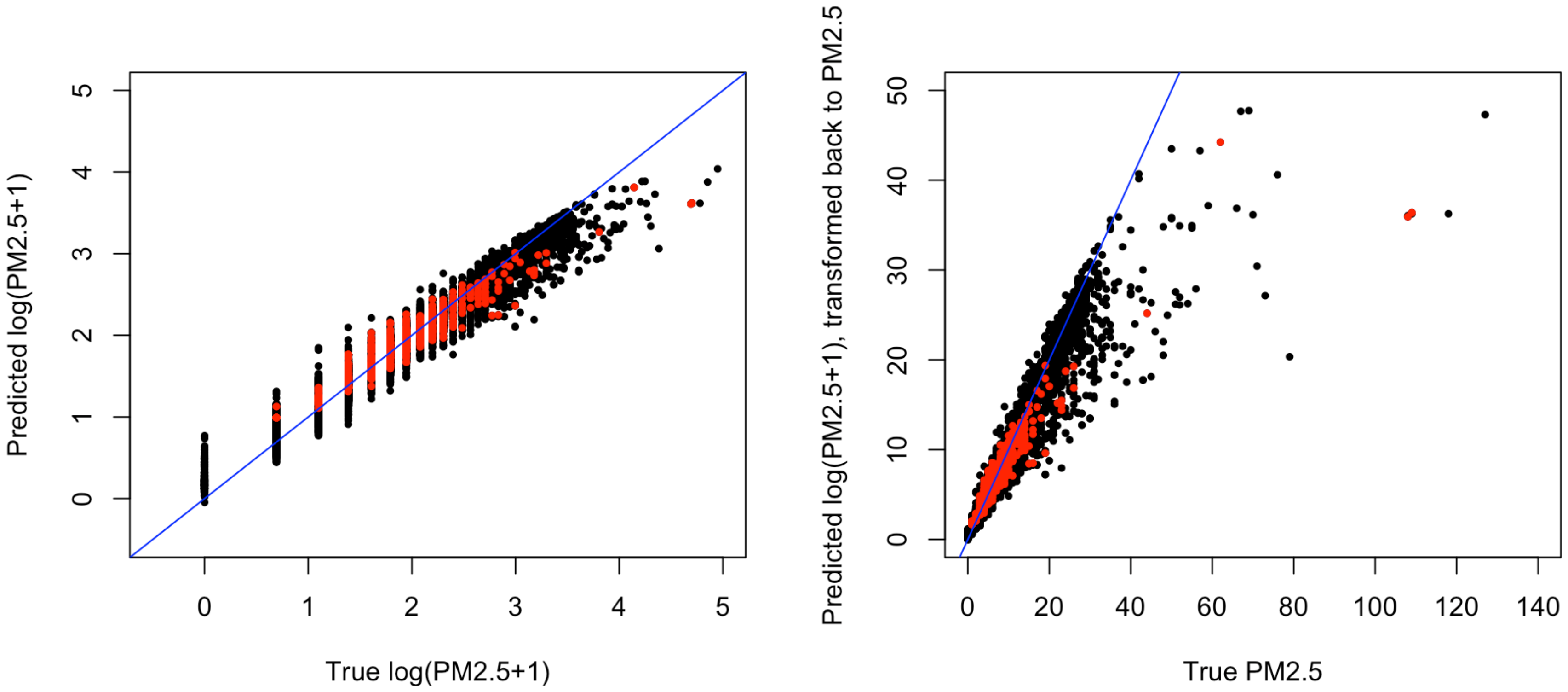
Mean of random field



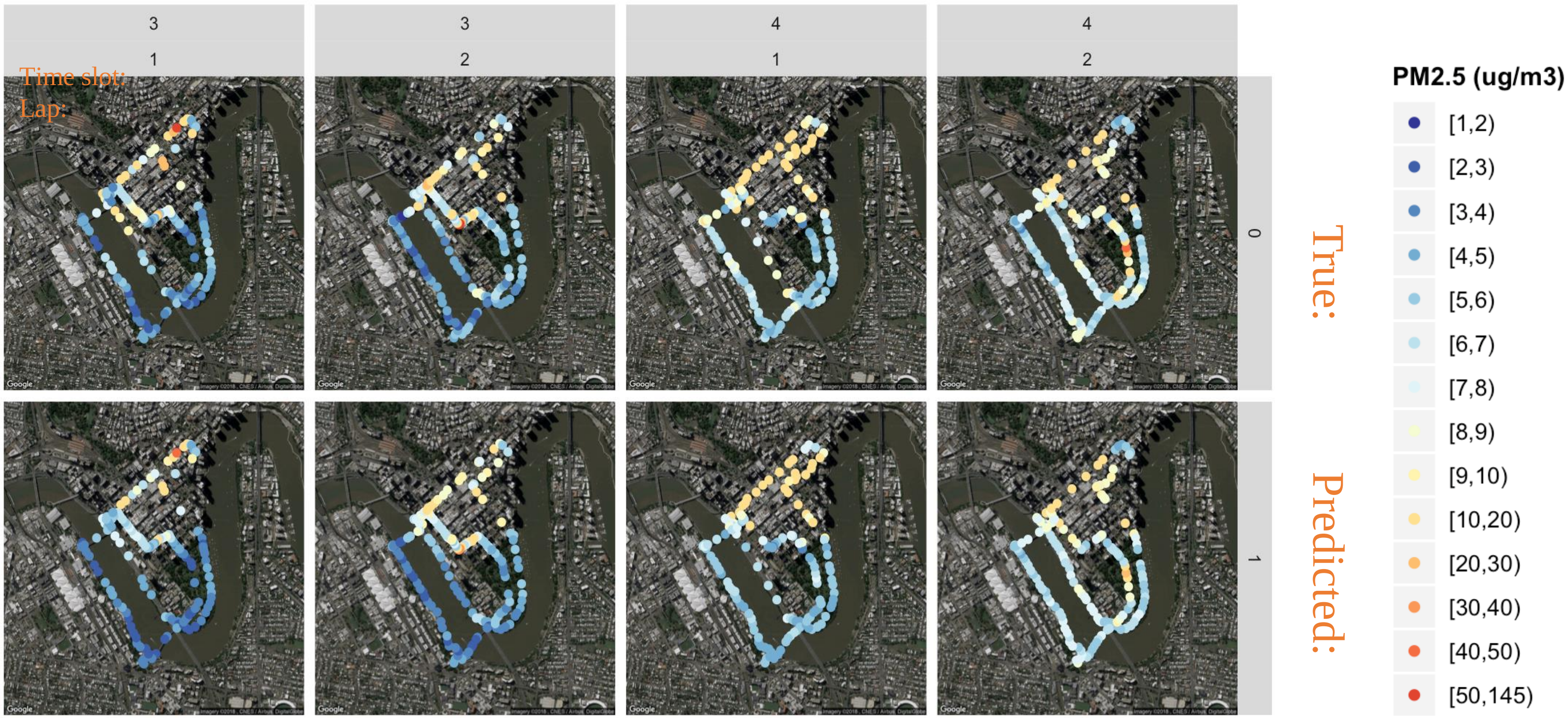
SD of random field

Predictive capability

Compare predicted y with true y



Predictive capability



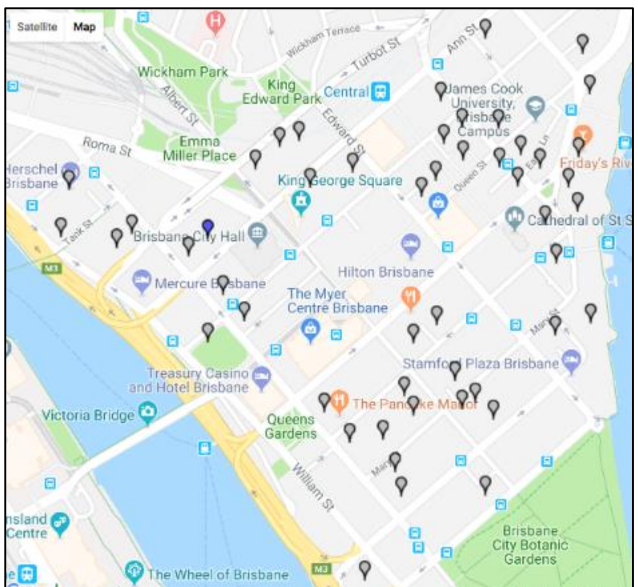
Could use the model to predict in any location, or along a route at regular intervals etc..

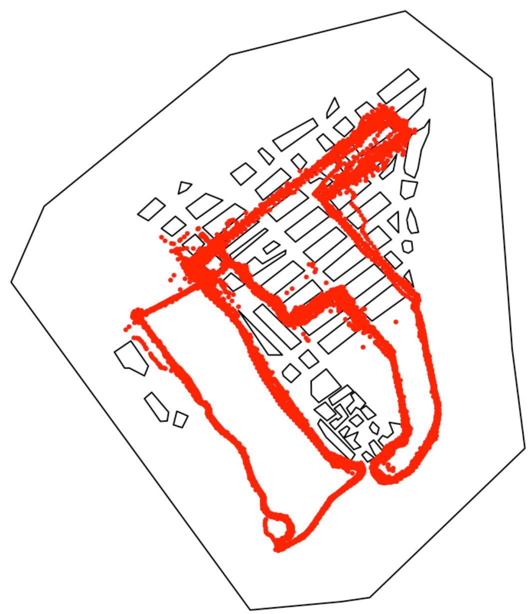
Barrier Model

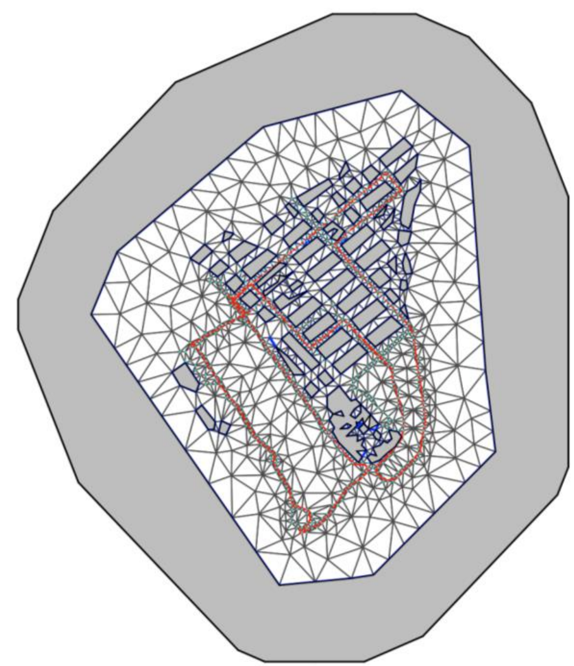
Haakon Bakka et al. (2018), **arXiv:1608.03787v2**

- In the barrier model "distance" is not the shortest distance, but rather a collection of all possible paths from one location to another; and the dependency between two points relies on all the paths that exists between them
- Implemented using INLA – same computational speed

Barrier created in ArcMap each block is a barrier

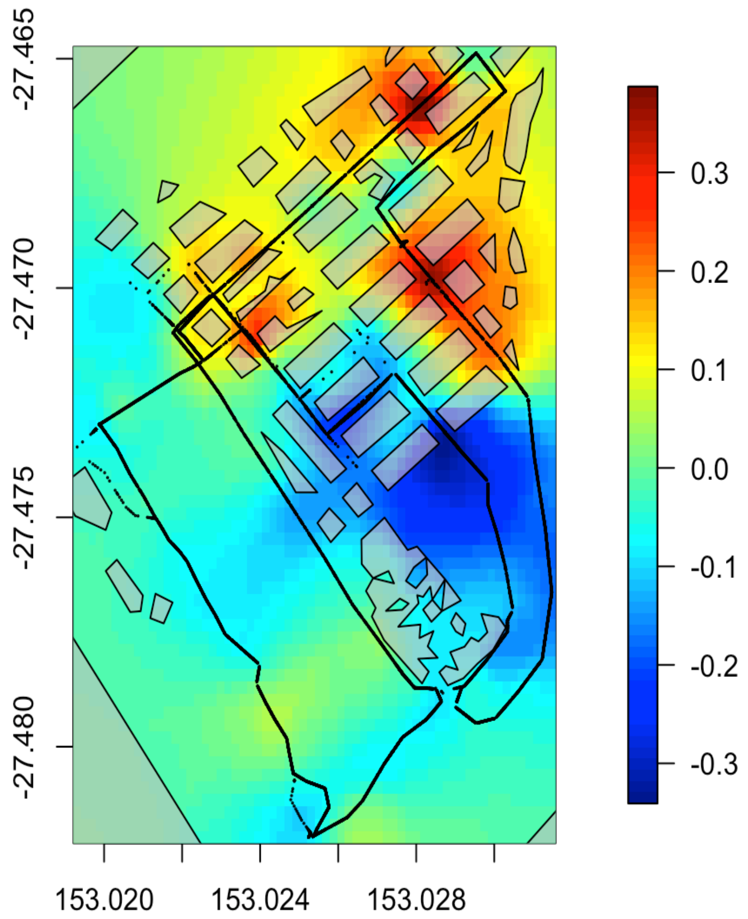




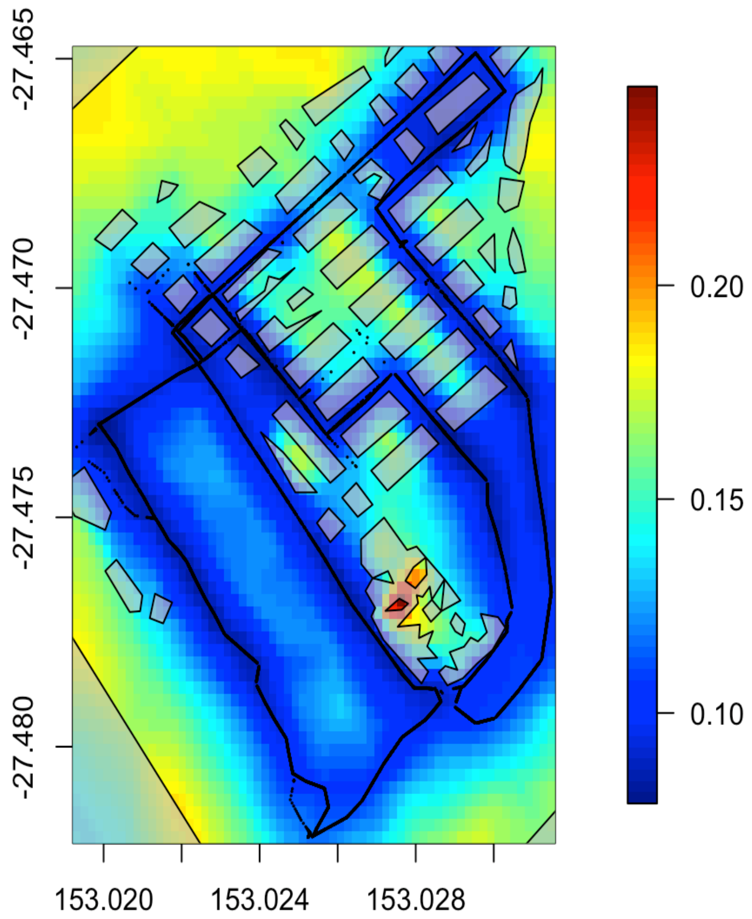


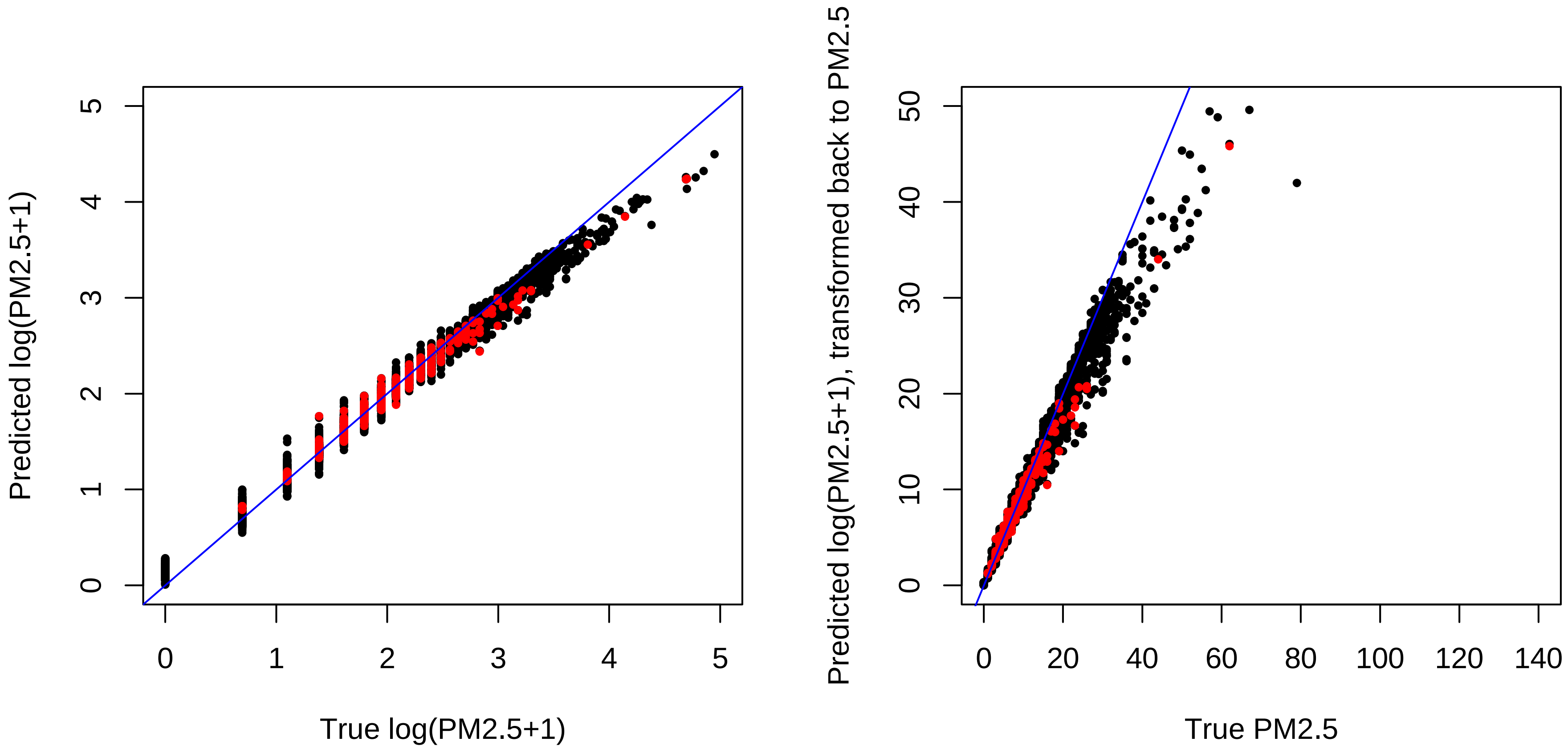
Results

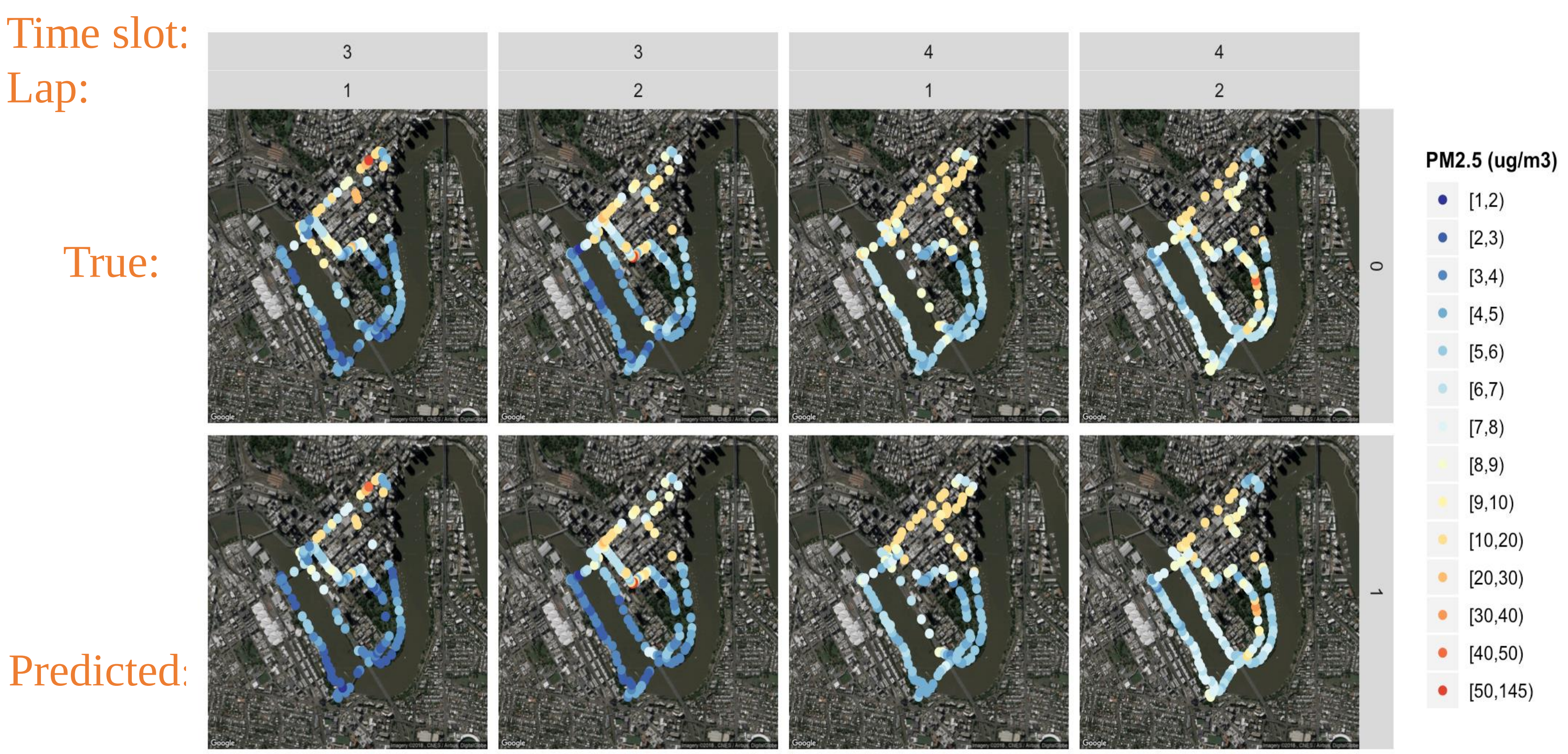
Mean of random field



SD of random field

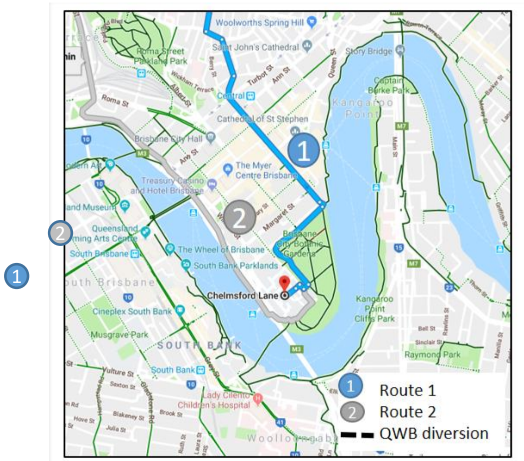


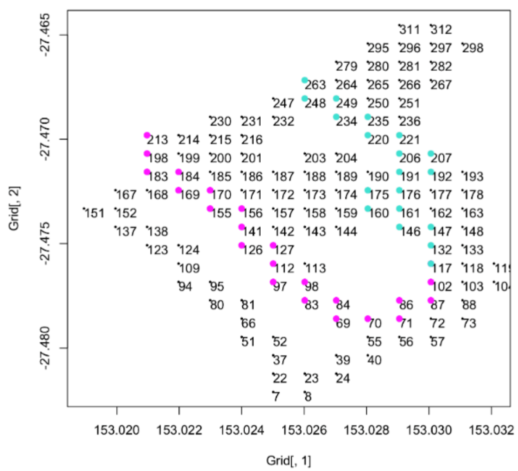

Predicted log(PM2.5+1)
True log(PM2.5+1)
Predicted log(PM2.5+1), transformed back to PM2.5
True PM2.5
0
1
2
3
4
5
0
10
20
30
40
50
0
20
40
60
80
100
120
140

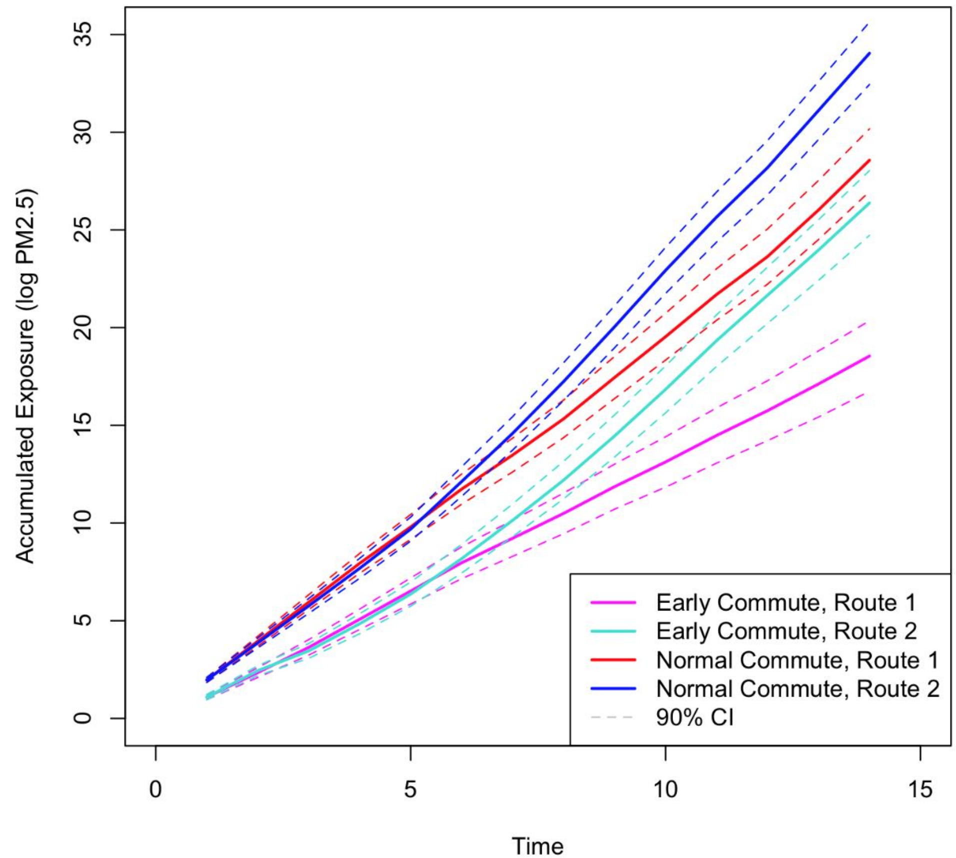


Time slot:
Lap:
True:
Predicted:
3
1
3
2
4
1
4
2
0
1
PM2.5 (ug/m3)
[1,2)
[2,3)
[3,4)
[4,5)
[5,6)
[6,7)
[7,8)
[8,9)
[9,10)
[10,20)
[20,30)
[30,40)
[40,50)
[50,145)

Accumulated Exposure along routes



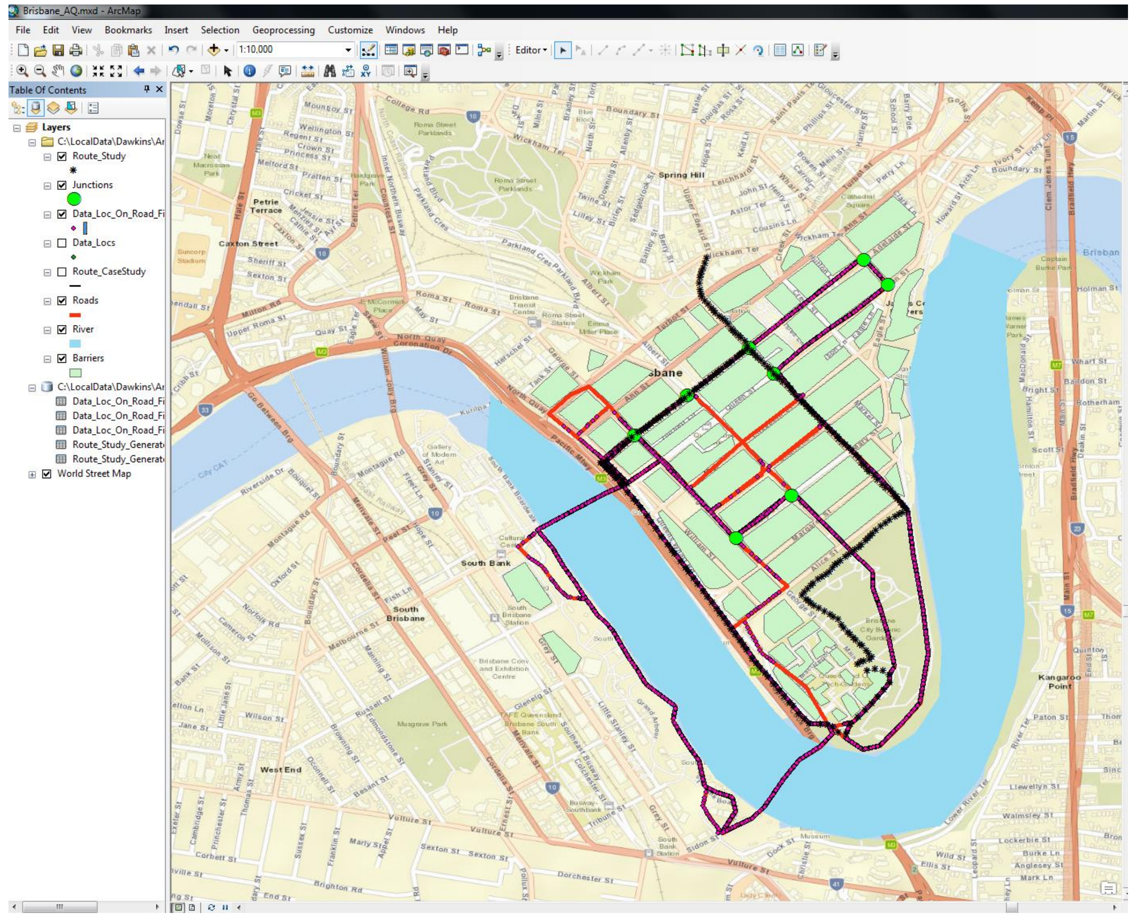


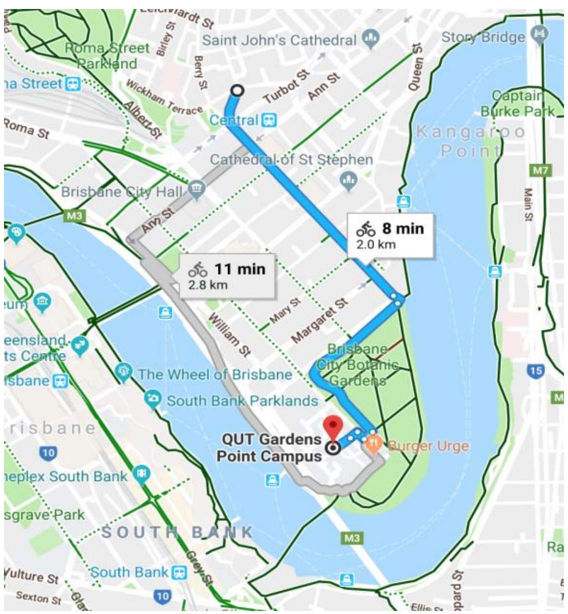


Making choices between routes

Using ArcGIS to create case study route and covariates

Blue route: faster but through more of CBD
Grey Route: slower and through less of CBD but on busy street



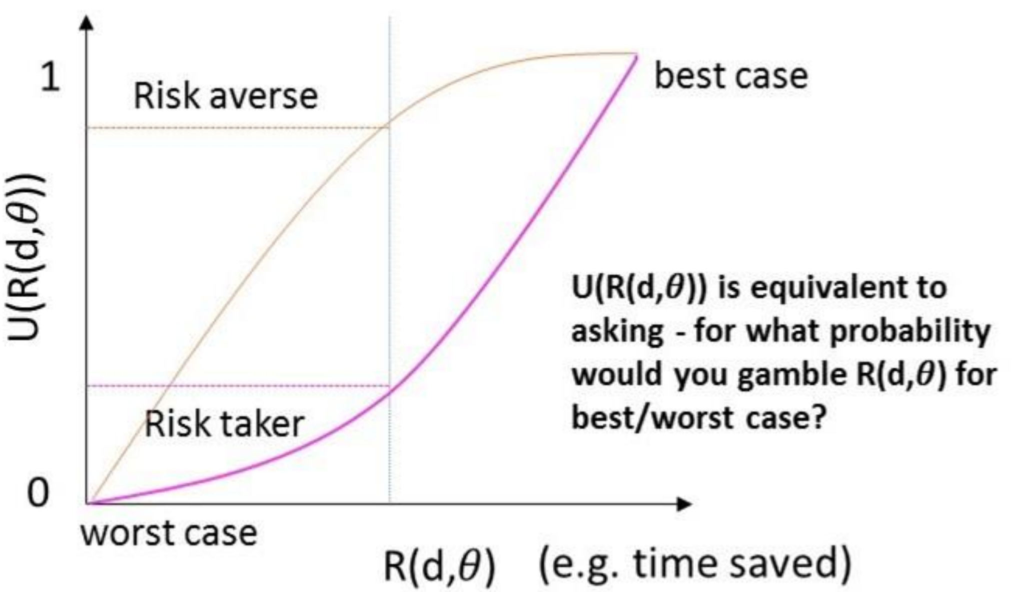


Decision Framework

- Find decision rule $d^*(y)$ that maximises the expected utility:

$$\bar{U}(d) = \int_{\theta \in \Theta} U(R(d,\theta))p(\theta|y)d\theta$$

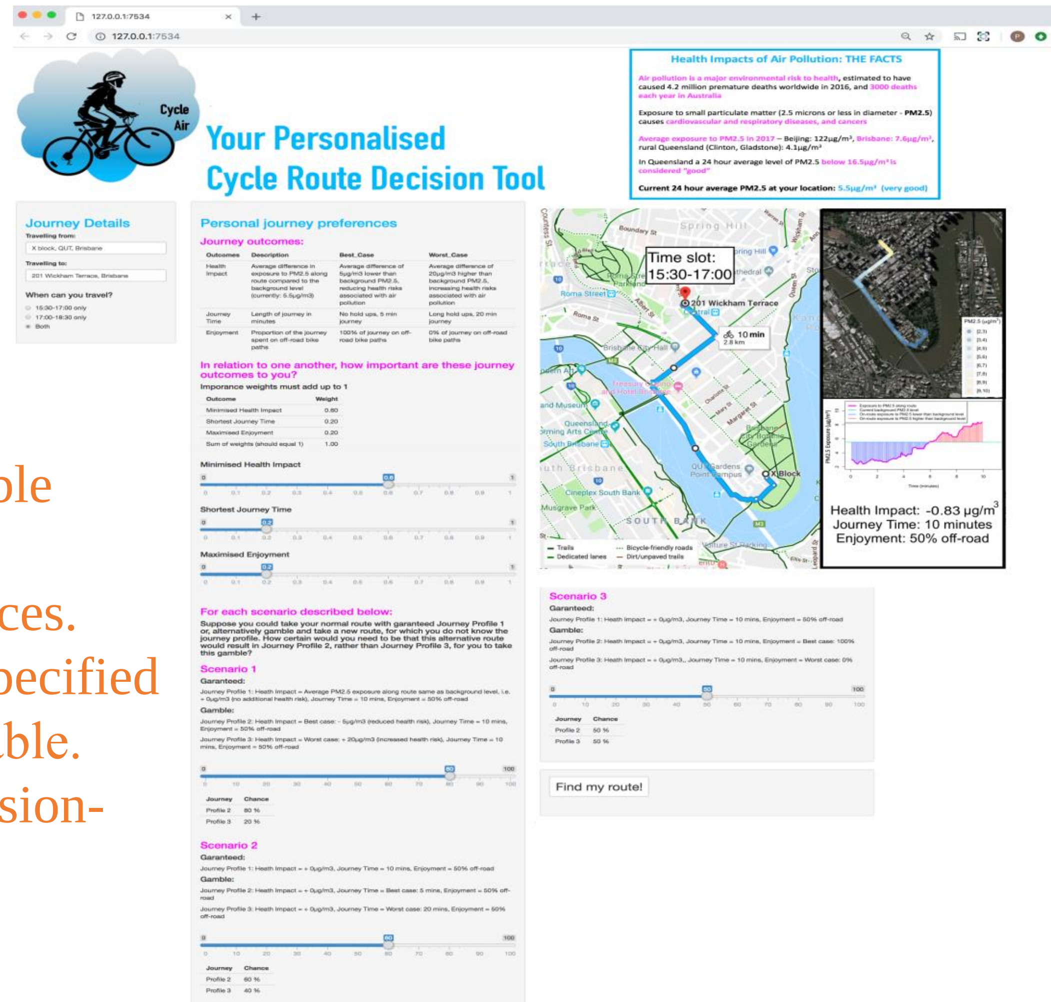
θ is the set outcomes/states of nature/consequences, y is the data, R is the pay off, U is the utility function, $p(\theta|y)$ is the joint density of the outcomes given the data



Decision Framework

- Outcomes: **health risk** and **time**
  - Our model gives us p(θ|y)
  - Elicit personal utility functions by asking about how risk averse someone is
  - Fit a log-log surface based on responses
  - Could also incorporate criterion weights based on a personal preferences:
    e.g. if pre-exiting health condition, *U* for health has more weight than *U* for time.

Eliciting the Utility function



1. Enter cycle journey details: where travel from/to, how flexible with time of travel
2. Elicit personal journey preferences.
3. Best and worst case scenarios specified within a “Journey Outcomes” table.
4. Use scenarios to determine decision-relevant attributes, for:
   k_1 = health impact
   k_2 = journey time
   k_3 = enjoyment

127.0.0.1:7534

Cycle Air

Your Personalised Cycle Route Decision Tool

Health Impacts of Air Pollution: THE FACTS

Air pollution is a major environmental risk to health, estimated to have caused 4.2 million premature deaths worldwide in 2016, and 3000 deaths each year in Australia

Exposure to small particulate matter (2.5 microns or less in diameter - **PM2.5**) causes cardiovascular and respiratory diseases, and cancers

Average exposure to PM2.5 in 2017 – Beijing: 122μg/m³, Brisbane: 7.6μg/m³, rural Queensland (Clinton, Gladstone): 4.1μg/m³

In Queensland a 24 hour average level of PM2.5 below 16.5μg/m³ is considered "good"

Current 24 hour average PM2.5 at your location: 5.5μg/m³ (very good)

Journey Details

Travelling from:

X block, QUT, Brisbane

Travelling to:

201 Wickham Terrace, Brisbane

When can you travel?

- 15:30-17:00 only
- 17:00-18:30 only
- Both

Personal journey preferences

Journey outcomes:

Outcomes	Description	Best_Case	Worst_Case
Health Impact	Average difference in exposure to PM2.5 along route compared to the background level (currently: 5.5μg/m3)	Average difference of 5μg/m3 lower than background PM2.5, reducing health risks associated with air pollution	Average difference of 20μg/m3 higher than background PM2.5, increasing health risks associated with air pollution
Journey Time	Length of journey in minutes	No hold ups, 5 min journey	Long hold ups, 20 min journey
Enjoyment	Proportion of the journey spent on off-road bike paths	100% of journey on off-road bike paths	0% of journey on off-road bike paths

In relation to one another, how important are these journey outcomes to you?

Imporance weights must add up to 1

Outcome	Weight
Minimised Health Impact	0.60
Shortest Journey Time	0.20
Maximised Enjoyment	0.20
Sum of weights (should equal 1)	1.00

Minimised Health Impact

0 — 0.6 — 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Shortest Journey Time

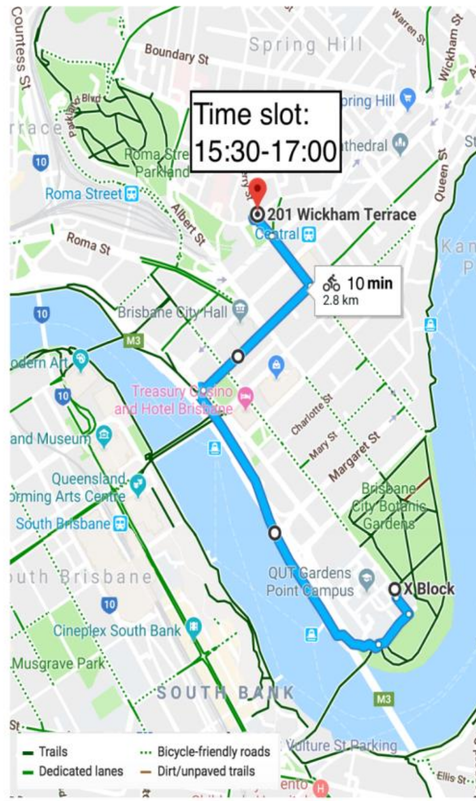
0 — 0.2 — 1

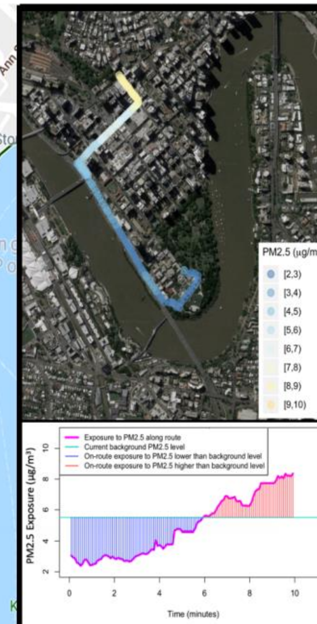
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Maximised Enjoyment

0 — 0.2 — 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1





Health Impact: -0.83 μg/m³
Journey Time: 10 minutes
Enjoyment: 50% off-road

Scenario 3

For each scenario described below:

Suppose you could take your normal route with garanteed Journey Profile 1 or, alternatively gamble and take a new route, for which you do not know the journey profile. How certain would you need to be that this alternative route would result in Journey Profile 2, rather than Journey Profile 3, for you to take this gamble?

Scenario 1

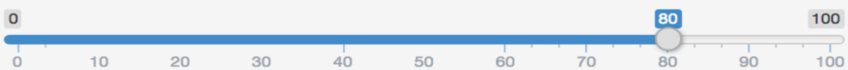
Garanteed:

Journey Profile 1: Heath Impact = Average PM2.5 exposure along route same as background level, i.e. + 0µg/m3 (no additional health risk), Journey Time = 10 mins, Enjoyment = 50% off-road

Gamble:

Journey Profile 2: Heath Impact = Best case: - 5µg/m3 (reduced health risk), Journey Time = 10 mins, Enjoyment = 50% off-road

Journey Profile 3: Heath Impact = Worst case: + 20µg/m3 (increased health risk), Journey Time = 10 mins, Enjoyment = 50% off-road



Journey	Chance
Profile 2	80 %
Profile 3	20 %

Scenario 2

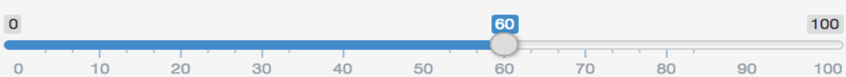
Garanteed:

Journey Profile 1: Heath Impact = + 0µg/m3, Journey Time = 10 mins, Enjoyment = 50% off-road

Gamble:

Journey Profile 2: Heath Impact = + 0µg/m3, Journey Time = Best case: 5 mins, Enjoyment = 50% off-road

Journey Profile 3: Heath Impact = + 0µg/m3, Journey Time = Worst case: 20 mins, Enjoyment = 50% off-road



Journey	Chance
Profile 2	60 %
Profile 3	40 %

Scenario 3

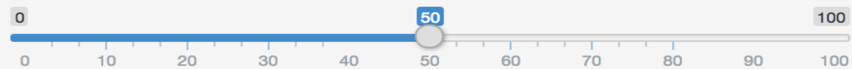
Garanteed:

Journey Profile 1: Heath Impact = + 0µg/m3, Journey Time = 10 mins, Enjoyment = 50% off-road

Gamble:

Journey Profile 2: Heath Impact = + 0µg/m3, Journey Time = 10 mins, Enjoyment = Best case: 100% off-road

Journey Profile 3: Heath Impact = + 0µg/m3,, Journey Time = 10 mins, Enjoyment = Worst case: 0% off-road



Journey	Chance
Profile 2	50 %
Profile 3	50 %

Find my route!

Case study

Journey A: Route 1, 15:30-17:00
Journey B: Route 1, 17:00-18:30
Journey C: Route 2, 15:30-17:00
Journey D: Route 2, 17:00-18:30

Journey	Health Impact	Journey Time	Enjoyment
A	-0.56 g/m3	10 mins	0.5 (prop. journey off-road)
B	0.44	11	0.45
C	-0.29	8.5	0.24
D	1.28	10	0.20

❖ If primary concern is health impact, then journey time: $(k_1, k_2, k_3) = (0.6; 0.2; 0.2)$

Expected utility: $(U_A, U_B, U_C, U_D) =$ $(0.716, 0.682, 0.686, 0.638)$

❖ If primary concern is journey time, then health impact $(k_1, k_2, k_3) = (0.3; 0.6; 0.1)$

Expected utility: (0.660, 0.623, 0.708, 0.614)

Scaling up

1. What is the role of priors in high dimensional regression?
2. What is the role of uncertainty and how do we represent it?
3. What about computational scalability?

1. What is the role of priors?

Wang X., Nott DJ, Drovandi CC, Mengersen K, Evans M (2018) Using history matching for prior choice. *Technometrics* 60, 445-460.

Design matrix X (n by p), $p > n$

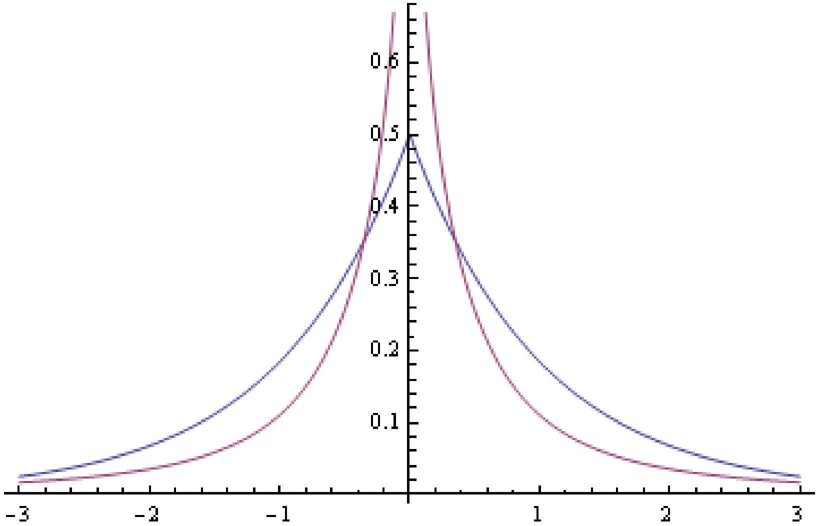
$$y = X\beta + \delta + \varepsilon \ , \quad \varepsilon \sim N(0, \sigma^2 I)$$

δ is a vector of mean shift parameters: want this to be sparse and allow for outliers in a small number of observations.

Priors: $\beta_j \sim \text{Horseshoe}(A_\beta)$

$\delta_j \sim \text{Horseshoe}(A_\delta)$

? How to choose priors for A_β , A_δ



Partially informative priors

Question:

What is a “reasonable” choice for a prior distribution from a parametric class?

(One) Answer:

Consider the *prior predictive distribution*

$$p(y) = \int p(\theta)p(y|\theta)d(\theta)$$

and use “history matching”

(complex infectious diseases, computer models, petroleum reservoir modelling, galaxy formation models, rainfall & climate models, etc)

History matching priors

Parameter of interest: θ

Class of prior distribution: $p(\theta|\lambda)$

Problem is to choose λ.

Treat the problem as one of model checking (for hypothetical data)

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_0^j that would be considered surprising if they were observed.

History matching priors

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

4. Compute $p_j(\lambda) = P(\log p(S^j|\lambda) \geq \log p(S_0^j|\lambda)$.

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

Repeat in *waves*, to obtain non-implausible priors.

Approximate Bayesian Computation

Calculate summary statistics $S(y)$ from observed data y

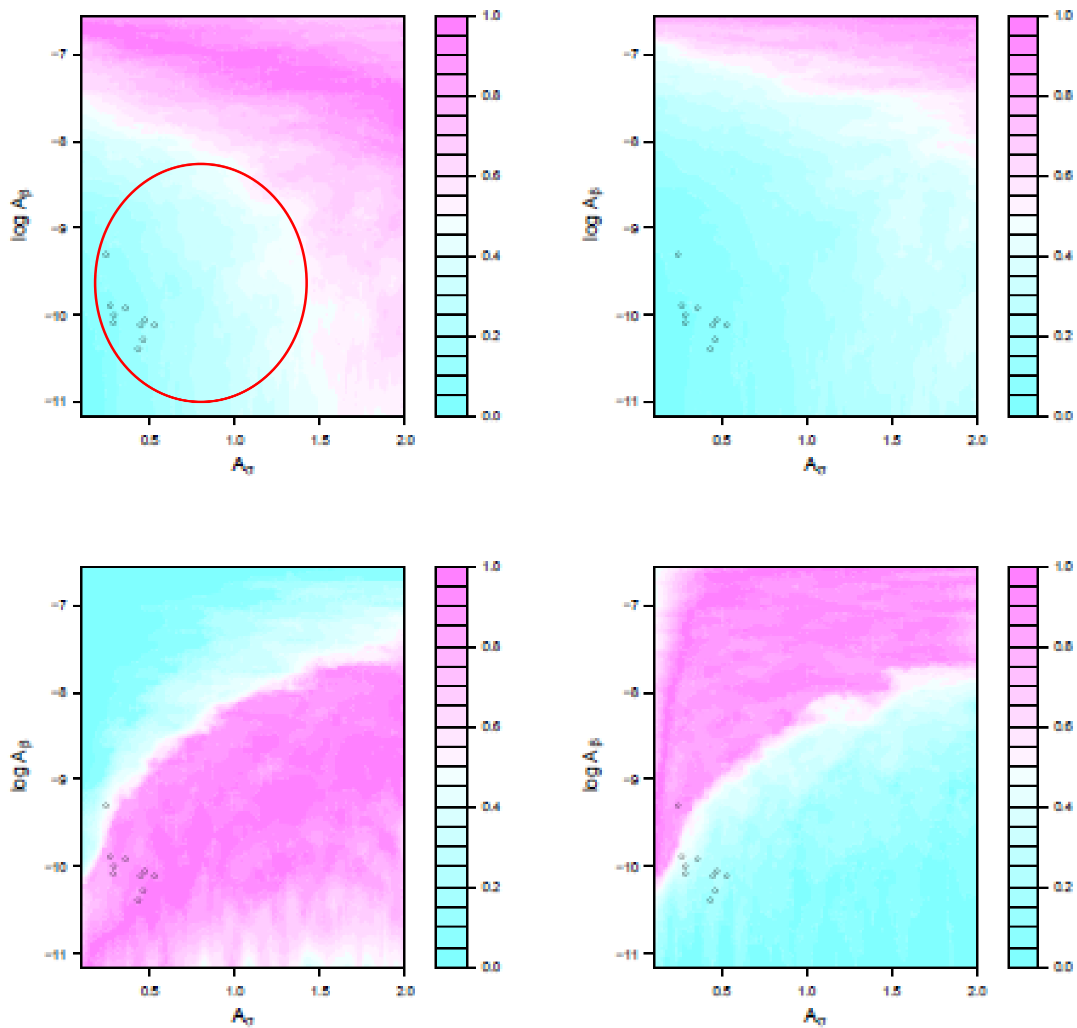
for t in 1:T **do**

1. Sample θ from the prior $p(\theta)$
2. Simulate pseudo-data y' from $p(y|\theta)$
3. Calculate $S(y')$
4. Accept θ if $\|S(y') - S(y)\| < \varepsilon$

end for

- No need for a likelihood
- Estimate complex models
- Simulate future unknowns (with uncertainty)

Example: sparse shrinkage regression



Surprising:

$S^1 = \log s^2 = \log 16$

$S^3 =$ refitted c-v $R^2 = 0.05$

Unsurprising:

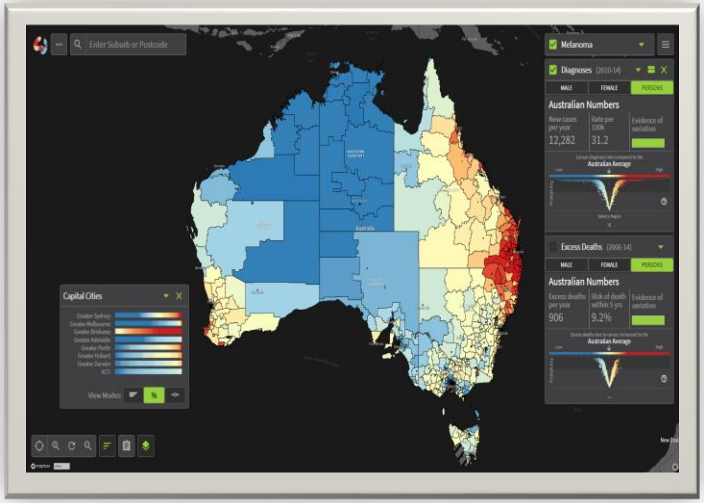
$S^2 = S^1 = \log 50$

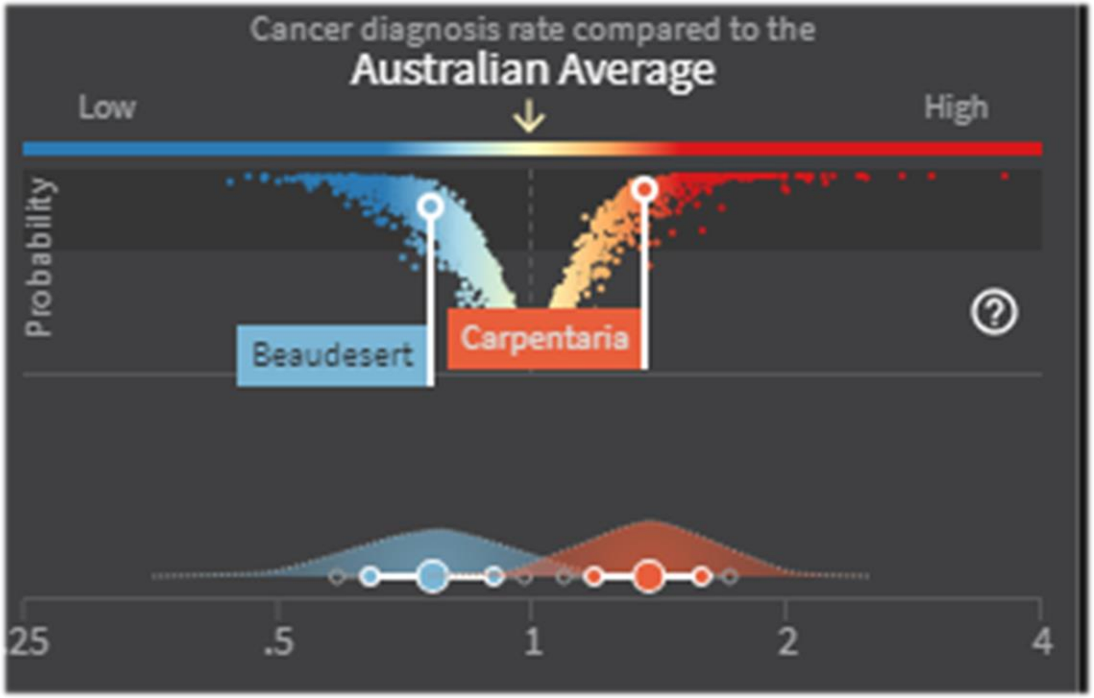
$S^4 = S^3 = 0.95$

2. What about uncertainty?

S. Cramb, E. Duncan, P. Baade, KM, *et al.*

http://atlas.cancer.org.au





3. What about scalable computation?

M. Sutton, E. Ebert, H. Xie, B. Liquet, C. Drovandi, D. Nott, KM, *et al.*

ABC

SMC

ABCSMC

ABCSMC^2

MCMC

INLA

VB

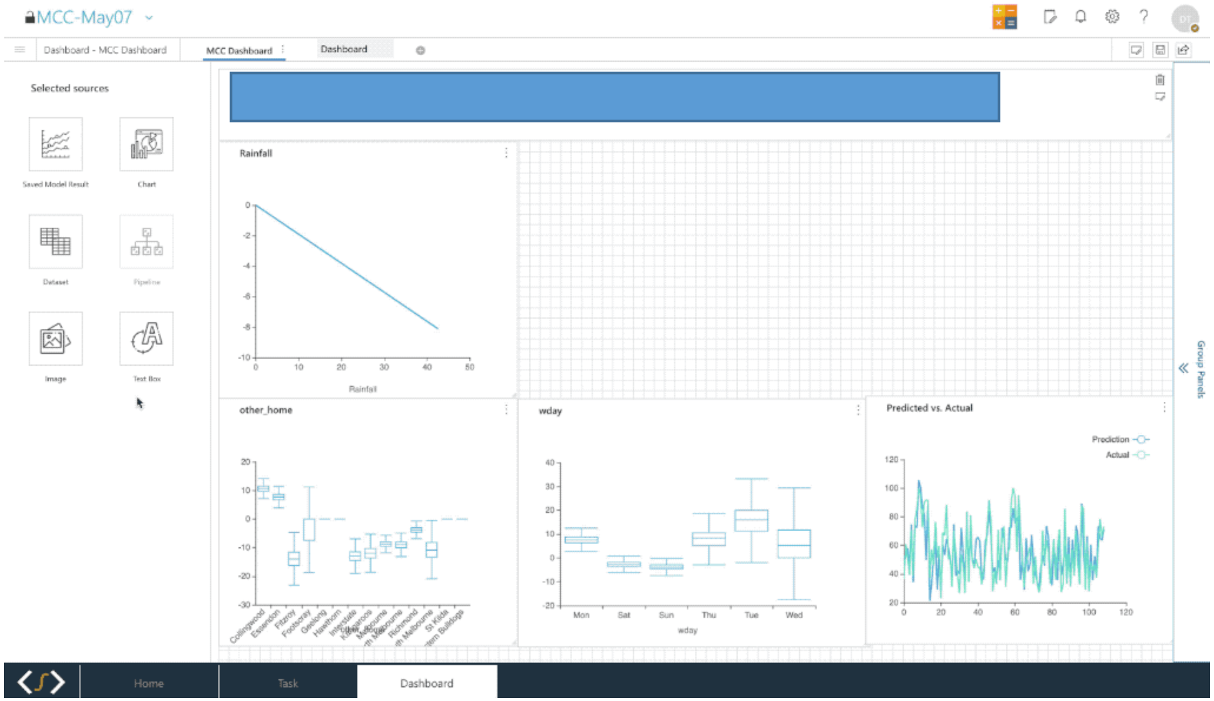
. . .

Liquet B.,
Mengersen K., Pettitt A.N.,
Sutton M. (2017) Bayesian variable selection regression of multivariate responses for group data.

Bayesian Analysis 4, 1039-1067.

AutoStat

https://autostat.com.au/



QUERY YOUR DATA

DASHBOARD AND COLLABORATION

DATA PRE-PROCESSING

CODE-FREE MODELLING

BUILD VISUALISATIONS

AUTOMATE MODEL SCHEDULING

GAIN NEW INSIGHTS

BAYESIAN OPTIMISATION

With acknowledgements to ACEMS and BRAG collaborators