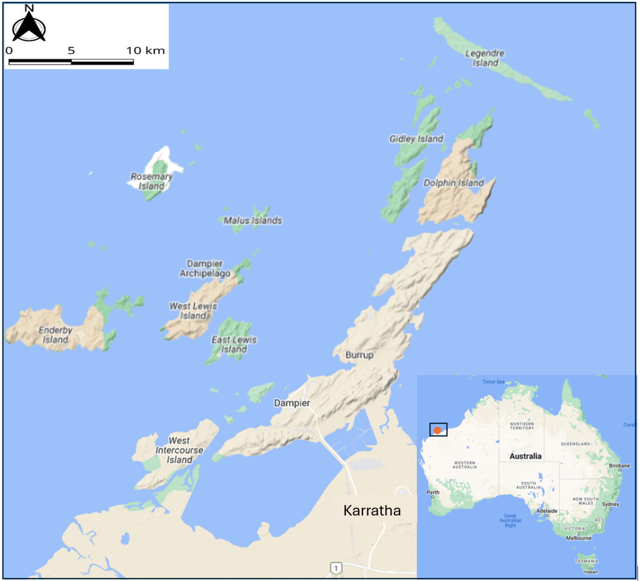
Spatial kernel smoothing with extreme outliers

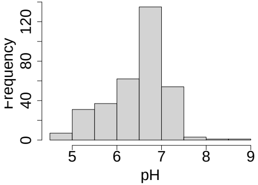
Dr Mohomed Abraj
Murujuga Rock Art Monitoring Program (MRAMP)
School of Population Health, Curtin University, Perth, WA.

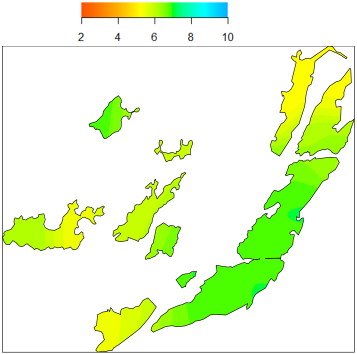
Australasian Applied Statistics Conference 2024

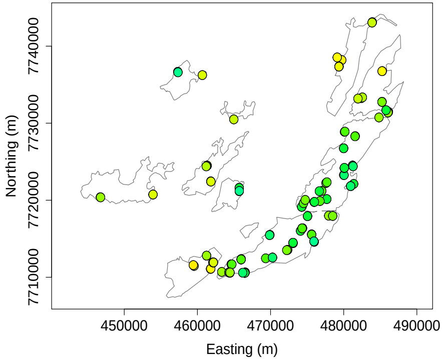
MRAMP study area



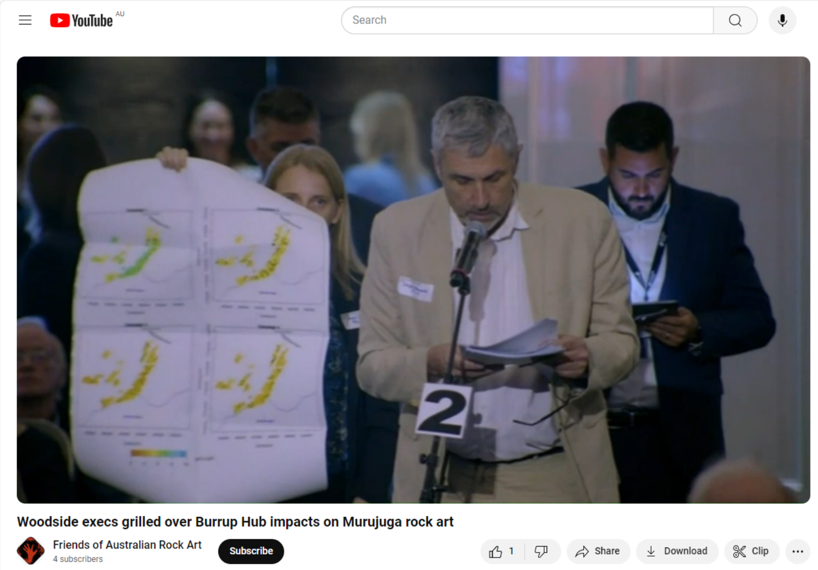
Sample size = 330



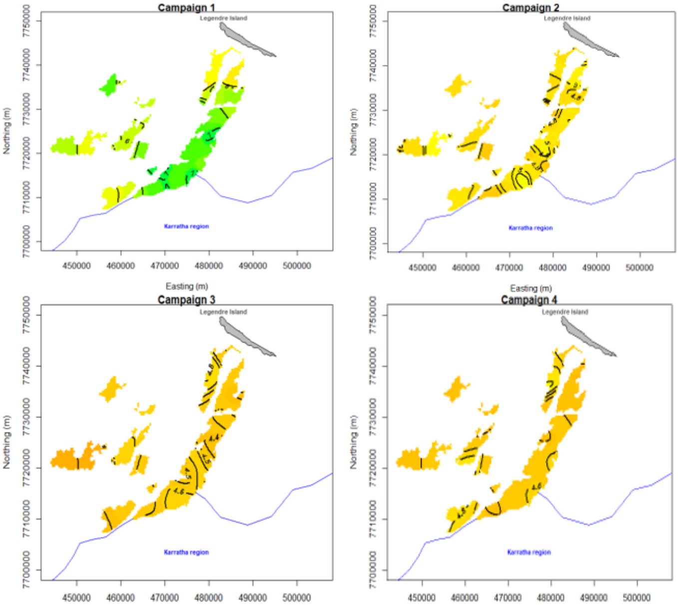




Woodside general meeting

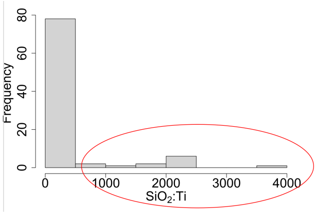


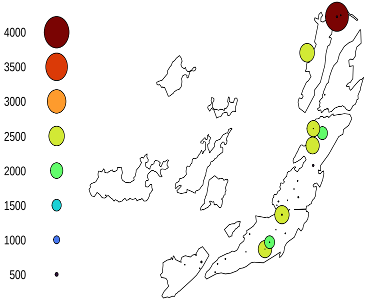
First year MRAMP report - spatial smoothing of pH



Data with extreme outliers - pXRF data

The **elemental composition** of rocks is evaluated with a portable X-ray fluorescence analyser (pXRF)





Spatial smoothing of elemental composition

- The **ratio** two elements is **more consistent** than the absolute pXRF measurement.
- Geologists often use **ratios** of two elements to characterise rock types.
- The **ratio of silica and titanium** (from pXRF data) was calculated, however, these ratios had a highly skewed distribution with **outliers**.
- The **Nadaraya-Watson (N-W) kernel smoother**, a non-parametric method, is computationally efficient method. However, N-W method is **not robust** with **extreme outliers**.
- Extreme outliers in the data can **distort** the smoothed surface and **mislead** the interpretation.

Spatial smoothing of elemental composition

- The **ratio** two elements is **more consistent** than the absolute pXRF measurement.
- Geologists often use **ratios** of two elements to characterise rock types.
- The **ratio of silica and titanium** (from pXRF data) was calculated, however, these ratios had a highly skewed distribution with **outliers**.
- The **Nadaraya-Watson (N-W) kernel smoother**, a non-parametric method, is computationally efficient method. However, N-W method is **not robust** with **extreme outliers**.
- Extreme outliers in the data can **distort** the smoothed surface and **mislead** the interpretation.

Spatial smoothing of elemental composition

- The **ratio** two elements is **more consistent** than the absolute pXRF measurement.
- Geologists often use **ratios** of two elements to characterise rock types.
- The **ratio of silica and titanium** (from pXRF data) was calculated, however, these ratios had a highly skewed distribution with **outliers**.
- The **Nadaraya-Watson (N-W) kernel smoother**, a non-parametric method, is computationally efficient method. However, N-W method is **not robust** with **extreme outliers**.
- Extreme outliers in the data can **distort** the smoothed surface and **mislead** the interpretation.

Spatial smoothing of elemental composition

- The **ratio** two elements is **more consistent** than the absolute pXRF measurement.
- Geologists often use **ratios** of two elements to characterise rock types.
- The **ratio of silica and titanium** (from pXRF data) was calculated, however, these ratios had a highly skewed distribution with **outliers**.
- The **Nadaraya-Watson (N-W) kernel smoother**, a non-parametric method, is computationally efficient method. However, N-W method is **not robust** with **extreme outliers**.
- Extreme outliers in the data can **distort** the smoothed surface and **mislead** the interpretation.

Spatial smoothing of elemental composition

- The **ratio** two elements is **more consistent** than the absolute pXRF measurement.
- Geologists often use **ratios** of two elements to characterise rock types.
- The **ratio of silica and titanium** (from pXRF data) was calculated, however, these ratios had a highly skewed distribution with **outliers**.
- The **Nadaraya-Watson (N-W) kernel smoother**, a non-parametric method, is computationally efficient method. However, N-W method is **not robust** with **extreme outliers**.
- Extreme outliers in the data can **distort** the smoothed surface and **mislead** the interpretation.

Nadaraya-Watson kernel smoothing

Let the data consist of the observed values $y_1, \ldots, y_n$ of some quantity y, observed at the sites $x_1, \ldots, x_n$ respectively.

$$y_i = Y(x_i) + \epsilon_i \tag{1}$$

where $Y(x), x \in R^2$ is a smooth function that is the target of investigation.

Then, the Nadaraya-Watson smoother is defined as

$$\widehat{Y}(x) = \frac{\sum_i y_i \, \kappa(x - x_i)}{\sum_i \kappa(x - x_i)}, \qquad x \in W \tag{2}$$

where $\kappa(x)$ is the smoothing kernel, a probability density on the two-dimensional plane.

R package **'spatstat'** by Adrian Baddeley

Nadaraya-Watson kernel smoothing

Let the data consist of the observed values $y_1, \ldots, y_n$ of some quantity y, observed at the sites $x_1, \ldots, x_n$ respectively.

$$y_i = Y(x_i) + \epsilon_i \tag{1}$$

where $Y(x), x \in R^2$ is a smooth function that is the target of investigation.

Then, the Nadaraya-Watson smoother is defined as

$$\widehat{Y}(x) = \frac{\sum_i y_i \ \kappa(x - x_i)}{\sum_i \kappa(x - x_i)}, \qquad x \in W \tag{2}$$

where $\kappa(x)$ is the smoothing kernel, a probability density on the two-dimensional plane.

R package **'spatstat'** by Adrian Baddeley

Winsorization (C P. Winsor[1])

Winsorization is a procedure in which the extremely high values of the data are replaced by **less extreme values**. Also, winsorized values are **more robust to outliers**.

For example data (N=20):

92, 19, 101, 58, 1053, 91, 26, 78, 10, 13, –40, 101, 86, 85, 15, 89, 89, 28, 5, 41

Data below the 5th percentile are –40 and 5.
Data above the 95th percentile are 101 and 1053.

A winsorized data would be:

92, 19, 101, 58, 101, 91, 26, 78, 10, 13, 5, 101, 86, 85, 15, 89, 89, 28, 5, 41

(–40 was replaced with 5 and 1053 was replaced with 101).

[1] N. J. Cox, *WINSOR: Stata module to Winsorize a variable*, Statistical Software Components, Boston College Department of Economics, Nov. 1998.

Winsorization (C P. Winsor[1])

Winsorization is a procedure in which the extremely high values of the data are replaced by **less extreme values**. Also, winsorized values are **more robust to outliers**.

For example data (N=20):

92, 19, 101, 58, 1053, 91, 26, 78, 10, 13, –40, 101, 86, 85, 15, 89, 89, 28, 5, 41

Data below the 5th percentile are –40 and 5.
Data above the 95th percentile are 101 and 1053.

A winsorized data would be:

92, 19, 101, 58, 101, 91, 26, 78, 10, 13, 5, 101, 86, 85, 15, 89, 89, 28, 5, 41

(–40 was replaced with 5 and 1053 was replaced with 101).

[1] N. J. Cox, *WINSOR: Stata module to Winsorize a variable*, Statistical Software Components, Boston College Department of Economics, Nov. 1998.

Winsorization (C P. Winsor[1])

Winsorization is a procedure in which the extremely high values of the data are replaced by **less extreme values**. Also, winsorized values are **more robust to outliers**.

For example data (N=20):

92, 19, 101, 58, 1053, 91, 26, 78, 10, 13, –40, 101, 86, 85, 15, 89, 89, 28, 5, 41

Data below the 5th percentile are –40 and 5.
Data above the 95th percentile are 101 and 1053.

A winsorized data would be:

92, 19, 101, 58, 101, 91, 26, 78, 10, 13, 5, 101, 86, 85, 15, 89, 89, 28, 5, 41

(–40 was replaced with 5 and 1053 was replaced with 101).

[1] N. J. Cox, *WINSOR: Stata module to Winsorize a variable*, Statistical Software Components, Boston College Department of Economics, Nov. 1998.

Winsorization

For values y_i, another winsorization procedure is as follows.

1. Calculate the inter-quartile range $IQR = Q_3 - Q_1$ of y_i.
2. Calculate upper and lower thresholds
$$U = Q_3 + c\,IQR, \qquad L = Q_1 - c\,IQR$$
where $c \geq 0$ is a chosen coefficient. A typical value is $c = 1.5$.
3. Calculate the extreme values within $[L, U]$,
$$\begin{aligned} y_U &= \max\{y_i : L \leq y_i \leq U\} \\ y_L &= \min\{y_i : L \leq y_i \leq U\} \end{aligned}$$
4. **Replace extreme values outside** $[L, U]$ with the nearest extreme inside $[L, U]$:
$$y_i^* = \begin{cases} y_L & \text{if } y_i < L \\ y_i & \text{if } L \leq y_i \leq U \\ y_U & \text{if } y_i > U \end{cases}$$

Winsorization

For values y_i, another winsorization procedure is as follows.

1. Calculate the inter-quartile range $IQR = Q_3 - Q_1$ of y_i.
2. Calculate upper and lower thresholds
$$U = Q_3 + c\,IQR, \qquad L = Q_1 - c\,IQR$$
where $c \geq 0$ is a chosen coefficient. A typical value is $c = 1.5$.
3. Calculate the extreme values within $[L, U]$,
$$\begin{aligned} y_U &= \max\{y_i : L \leq y_i \leq U\} \\ y_L &= \min\{y_i : L \leq y_i \leq U\} \end{aligned}$$
4. **Replace extreme values outside** $[L, U]$ with the nearest extreme inside $[L, U]$:
$$y_i^* = \begin{cases} y_L & \text{if } y_i < L \\ y_i & \text{if } L \leq y_i \leq U \\ y_U & \text{if } y_i > U \end{cases}$$

Winsorization

For values y_i, another winsorization procedure is as follows.

1. Calculate the inter-quartile range $IQR = Q_3 - Q_1$ of y_i.
2. Calculate upper and lower thresholds

   $$U = Q_3 + c\,IQR, \qquad L = Q_1 - c\,IQR$$

   where $c \geq 0$ is a chosen coefficient. A typical value is $c = 1.5$.
3. Calculate the extreme values within $[L, U]$,

   $$\begin{aligned} y_U &= \max\{y_i : L \leq y_i \leq U\} \\ y_L &= \min\{y_i : L \leq y_i \leq U\} \end{aligned}$$

4. **Replace extreme values outside** $[L, U]$ with the nearest extreme inside $[L, U]$:

   $$y_i^* = \begin{cases} y_L & \text{if } y_i < L \\ y_i & \text{if } L \leq y_i \leq U \\ y_U & \text{if } y_i > U \end{cases}$$

Winsorization

For values y_i, another winsorization procedure is as follows.

1. Calculate the inter-quartile range $IQR = Q_3 - Q_1$ of y_i.
2. Calculate upper and lower thresholds

   $$U = Q_3 + c\, IQR, \qquad L = Q_1 - c\, IQR$$

   where $c \geq 0$ is a chosen coefficient. A typical value is $c = 1.5$.
3. Calculate the extreme values within $[L, U]$,

   $$\begin{aligned} y_U &= \max\{y_i : L \leq y_i \leq U\} \\ y_L &= \min\{y_i : L \leq y_i \leq U\} \end{aligned}$$

4. **Replace extreme values outside** $[L, U]$ with the nearest extreme inside $[L, U]$:

   $$y_i^* = \begin{cases} y_L & \text{if } y_i < L \\ y_i & \text{if } L \leq y_i \leq U \\ y_U & \text{if } y_i > U \end{cases}$$

Winsorization

For values y_i, another winsorization procedure is as follows.

1. Calculate the inter-quartile range $IQR = Q_3 - Q_1$ of y_i.
2. Calculate upper and lower thresholds

   $$U = Q_3 + c\,IQR, \qquad L = Q_1 - c\,IQR$$

   where $c \geq 0$ is a chosen coefficient. A typical value is $c = 1.5$.
3. Calculate the extreme values within $[L, U]$,

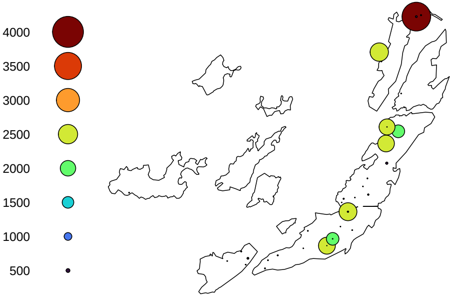
   $$\begin{aligned} y_U &= \max\{y_i : L \leq y_i \leq U\} \\ y_L &= \min\{y_i : L \leq y_i \leq U\} \end{aligned}$$

4. **Replace extreme values outside** $[L, U]$ with the nearest extreme inside $[L, U]$:

   $$y_i^* = \begin{cases} y_L & \text{if } y_i < L \\ y_i & \text{if } L \leq y_i \leq U \\ y_U & \text{if } y_i > U \end{cases}$$

Ratio of SiO_2 and Ti in Granophyre

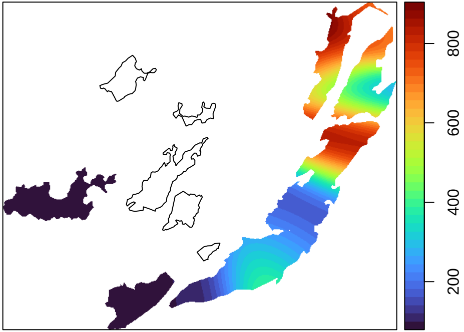
Raw data



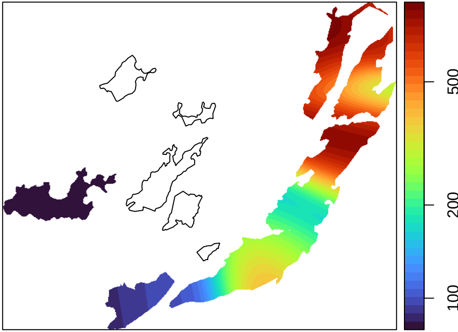
Winsorized data

Spatial smoothing for ratio of SiO_2 and Ti

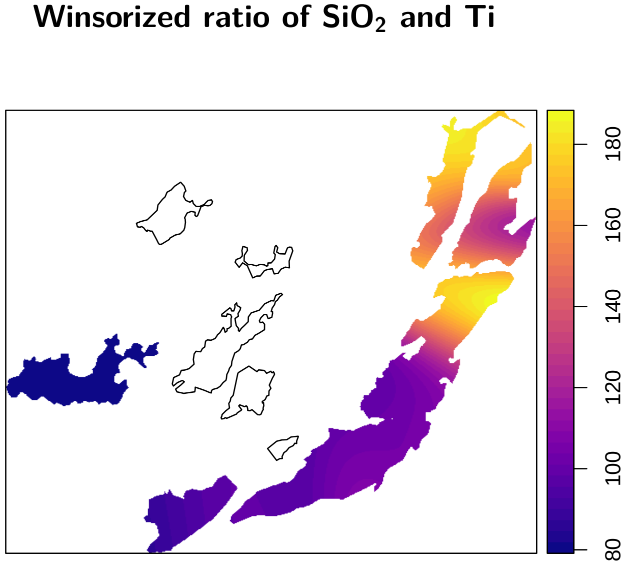
Raw data



Raw data (log scale)



Spatial smoothing of winsorized ratio



Cross-validation

Cross validation. Minimum error measurements are in [bold] [2].

Ratio of SiO_2 and Ti	MAE	RMSE	BIAS
Raw data	387.65	686.54	-4.55
Raw data (log)	283.51	692.55	188.45
Winsorized data	**52.50**	**65.68**	**-0.71**

Thus, winsorization is an **efficient method** to reduce the effect of **spurious outliers** in N-W spatial kernel smoothing.

[2] MAE (mean absolute error), RMSE (root mean square error)

Acknowledgement

I would like to express my gratitude to all those who have supported and contributed to this research. I particularly grateful to Professor Benjamin Mullins, Distinguished Professor Adrian Baddeley and Distinguished Professor Noel Cressie for their invaluable guidance. I also thank Dr Rebecca O'Leary, Dr Stephanie Hogg, Dr Suman Rakshit for their invaluable advice. This work was supported by MAC (Murujuga Aboriginal Corporation), and DWER (Department of Water and Environmental Regulation). I also thank Dr Tommaso Tacchetto, Professor Katy Evans, Glen Aubrey, Kasziem Bin-Sali, and Professor Pete Kinny, for their collaboration and assistance.