Statistical Inference and Data Analysis for SPM

UKSPM

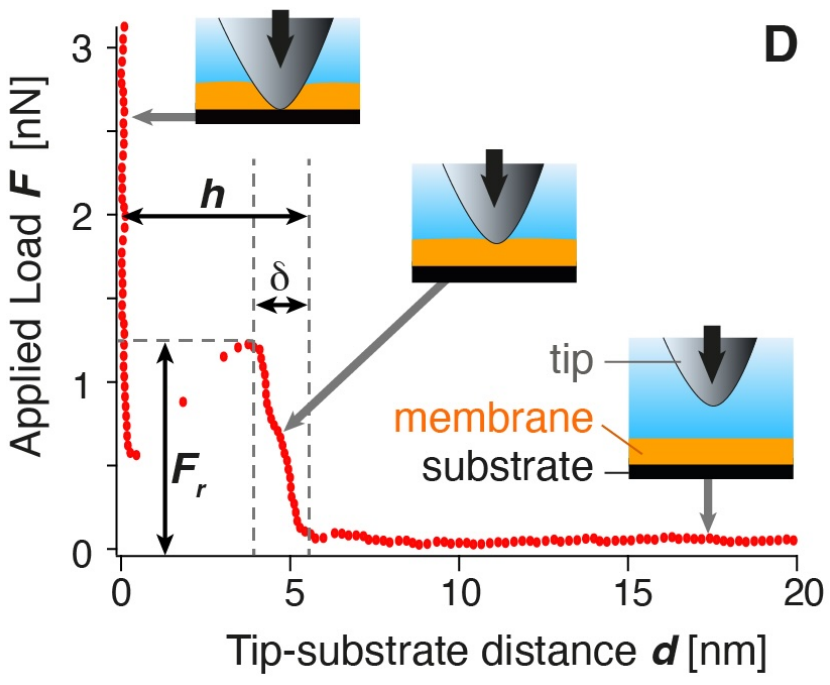
Andrew Iskauskas

Durham University

25th March, 2024

Talk Structure

Motivation



- AFM experiments generate large quantities of (potentially high-dimensional) data, under various assumptions, from which we want to perform inference.
- We want to say meaningful things about properties at each site: we *must* reduce the data to a manageable form.

Data courtesy of K. Voïtchovsky

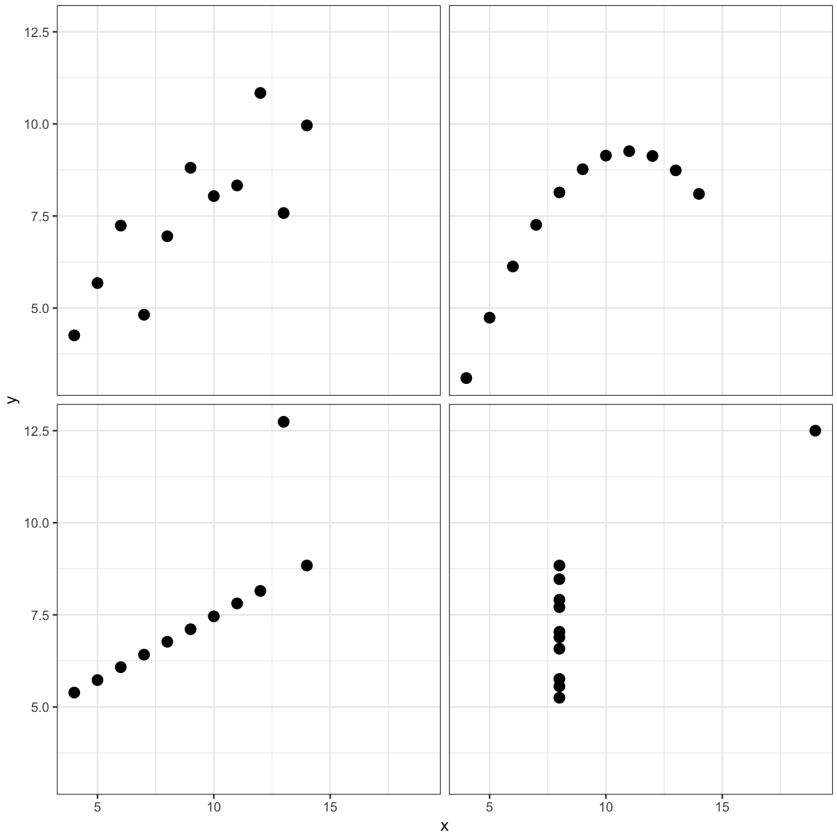
Summary Statistics?

What's wrong with taking data, calculating summary statistics, and moving on with our lives?
What if we had $n = 11$ observations from a bivariate dataset (x, y), with the following sample summary statistics (correct to 2dp):

E[x]	Var[x]	E[y]	Var[y]
9	11	7.5	4.13
Corr[x, y]	**$y = mx + c$: c**	**$y = mx + c$: m**	**R^2 value**
0.82	3	0.5	0.67

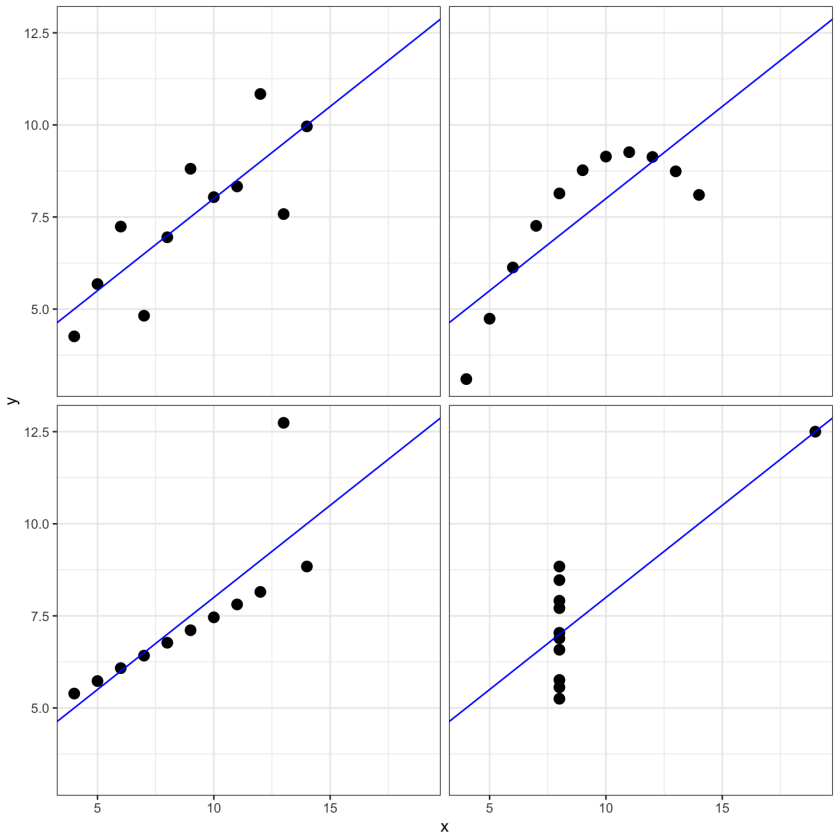
Can we say anything about the structure/distribution of the datapoints themselves?

Spot the Dataset



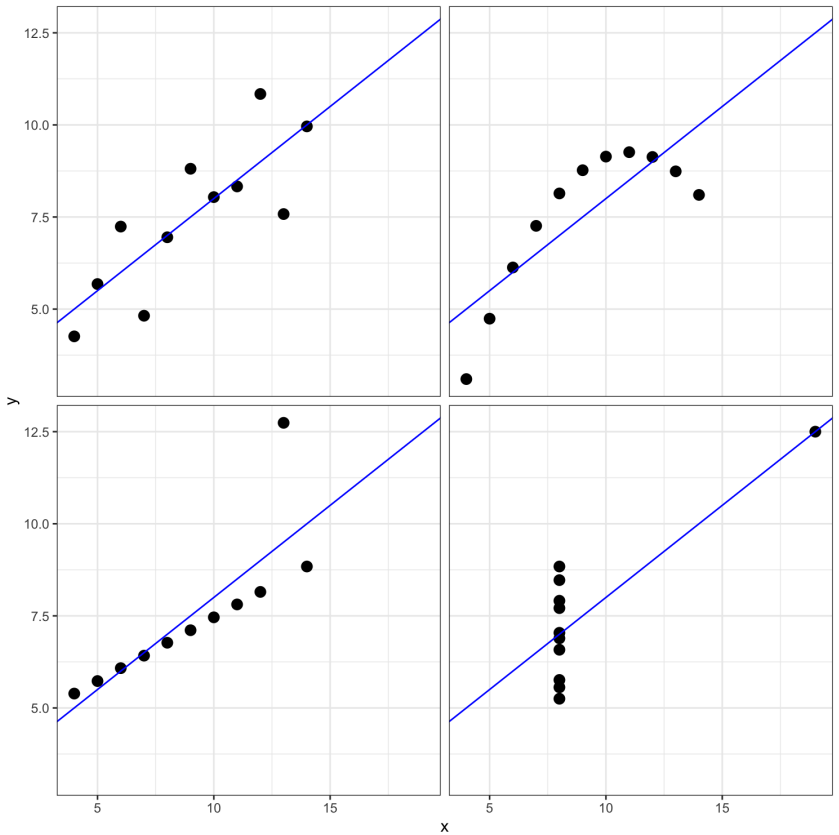
E[x]	Var[x]
9	11
E[y]	Var[y]
7.5	4.125
Corr[x, y]	c
0.816	3
m	R^2 value
0.5	0.67

Spot the Dataset



E[x]	Var[x]
9	11
E[y]	**Var[y]**
7.5	4.125
Corr[x, y]	**c**
0.816	3
m	**R^2 value**
0.5	0.67

Spot the Dataset

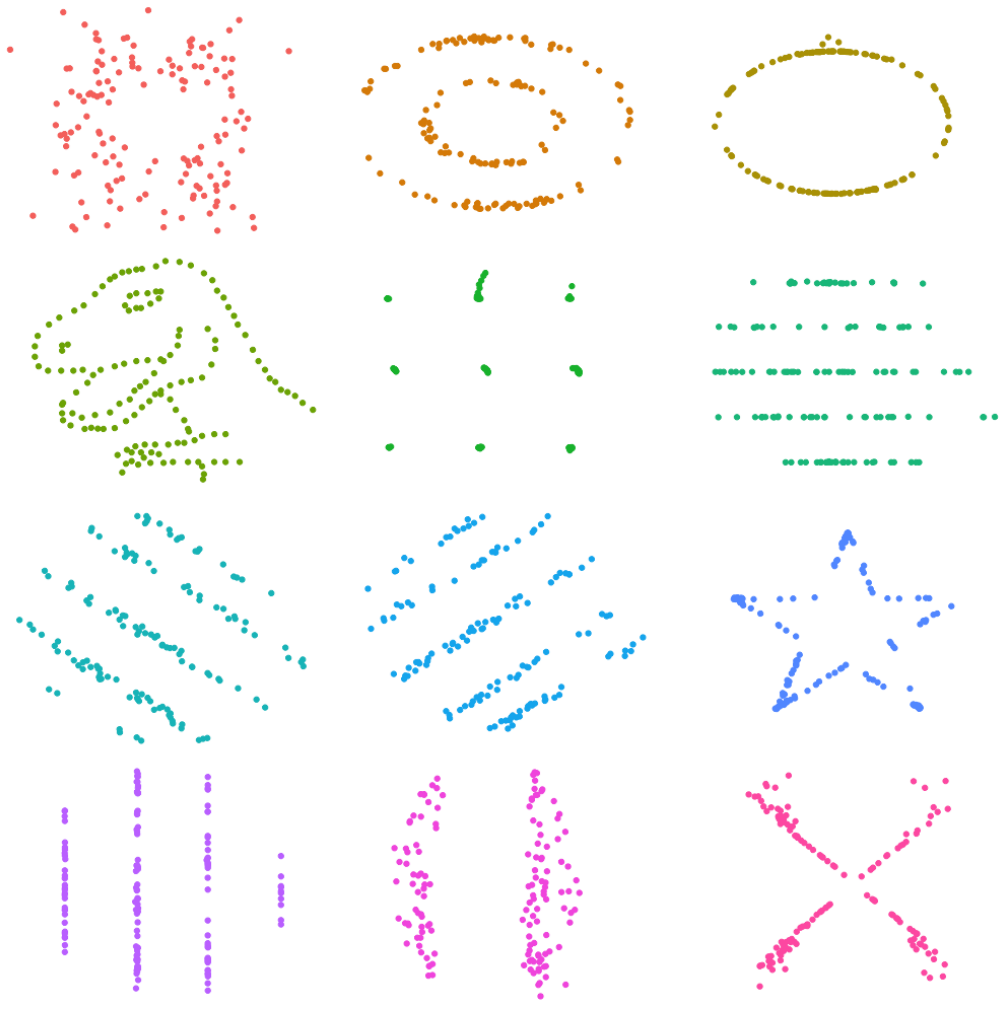


E[x]	Var[x]
9	11
E[y]	Var[y]
7.5	4.125
Corr[x, y]	c
0.816	3
m	R^2 value
0.5	0.67

All four datasets have the relevant summary statistics!

Driving the Point Home with a Dinosaur

The "Datasaurus Dozen": same principle, more complex.



Propagation of (Human) Error

The Anscombe set is a nice example of the pitfalls of assumptions, but without any obvious physical application.

However, suppose somebody asked you:

- I have a new observation of $x = 15$: what's the corresponding y?
- I observed $x = 15$ and $y = 7.5$: is this an abnormal value to observe?
- How confident should I be of the above predictions?

Our considerations matter a lot in those circumstances!

Framework for Statistical Inference

The Anscombe Quartet highlights three key steps we might wish to follow.

1. Considering and potentially removing spurious or misleading data alerts us to outliers, high leverage points, etc
2. Visualising the data reinforces/rebuffs any pre-existing assumptions we have about the data
3. Critical consideration of statistical assumptions (normality? Linearity?) with the help of the above alerts us to appropriate statistical tests for our data.

Once done, we can be confident that the quantitative results we obtain are *reliable*, *robust*, and *relevant*.

Outline

Using some real data, we will highlight the steps that we might take to ensure robust prediction and inference:

- Establish our modelling assumptions
- Investigate the data
- Determine our summary values, *and determine our uncertainty structure about them*
- Investigate outliers and/or abnormal values
- Consider data imputation where relevant
- Design a statistical test.

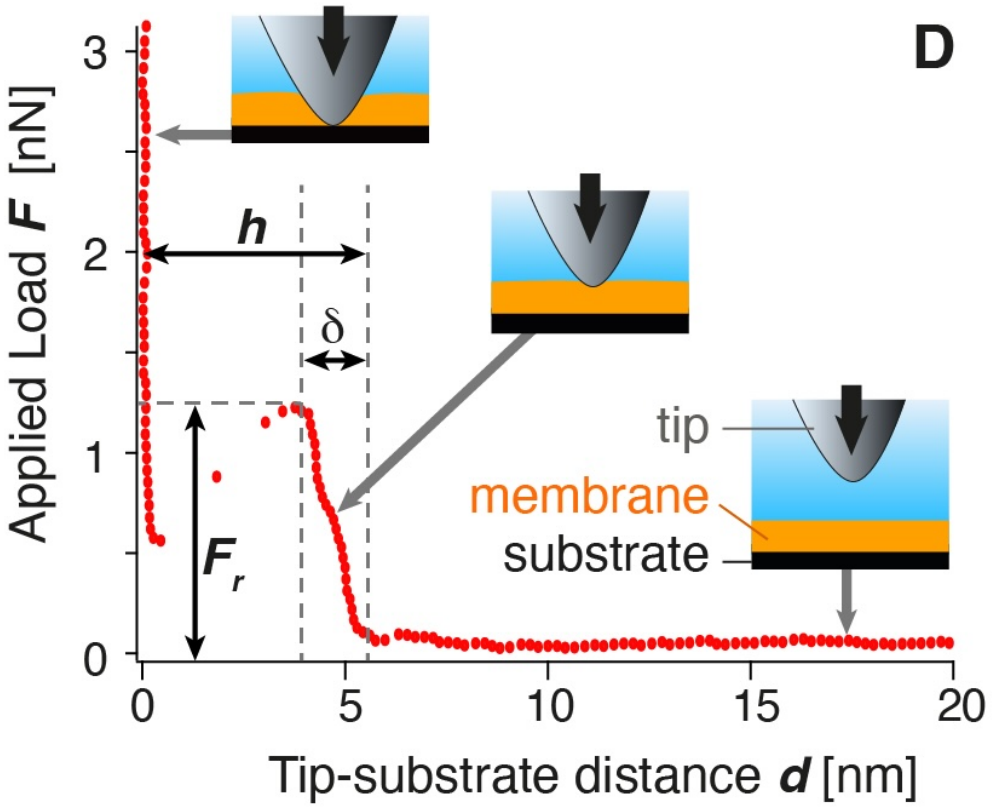
Caveat Emptor!

I will present an approach to dealing with the data – it is not *the* approach.

No amount of statistical machinery replaces expert judgment; if you disagree with choices I make, that's perfectly reasonable (and unsurprising)!

The point is that we *clearly define* the choices we make, so that we know exactly the context that we make final statements and we can defend them robustly.

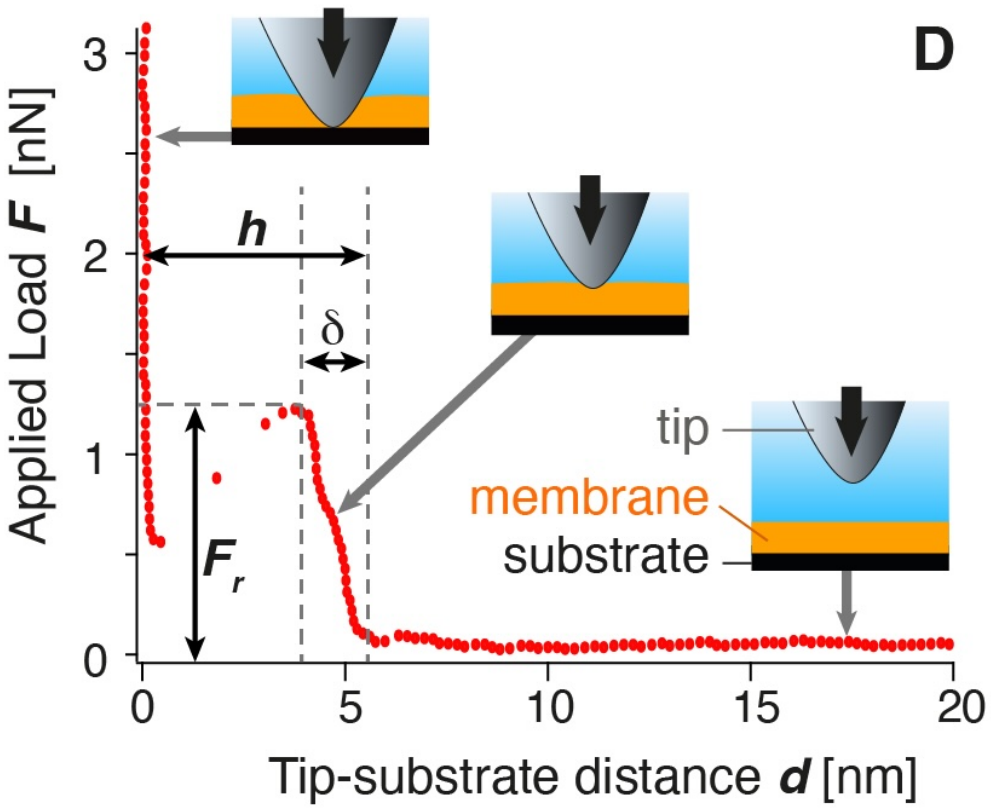
Data Structure



At each experiment site, we have a (noisy) curve, from which we want to determine:

- The point at which the probe touches the membrane;
- The rupture point and maximal force at rupture;
- The membrane thickness.

Data Structure



Even by-eye, the exact values of h and δ are not straightforward to elicit; how do we deal with the uncertainty this produces? How do we generalise this to avoid having to do this at every experiment site?

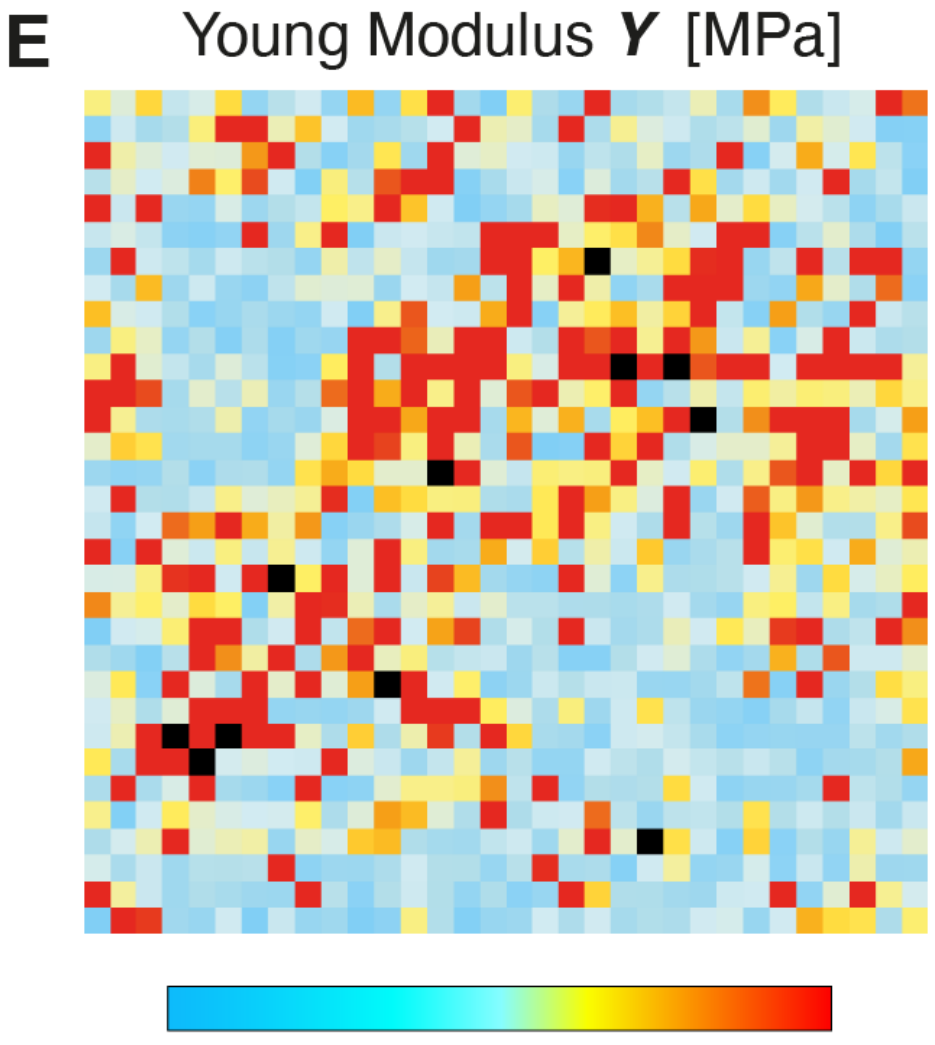
The Experimental Questions

We need to identify the following:

- The point at which the probe touches the membrane, x_{d_1};
- The rupture point, x_{d_2};
- The force $F_r(x_{d_2})$ at the rupture point.

We assume a common behaviour of the plots: flat until d_1, approximately linear to d_2, then a precipitous drop until $x = 0$.

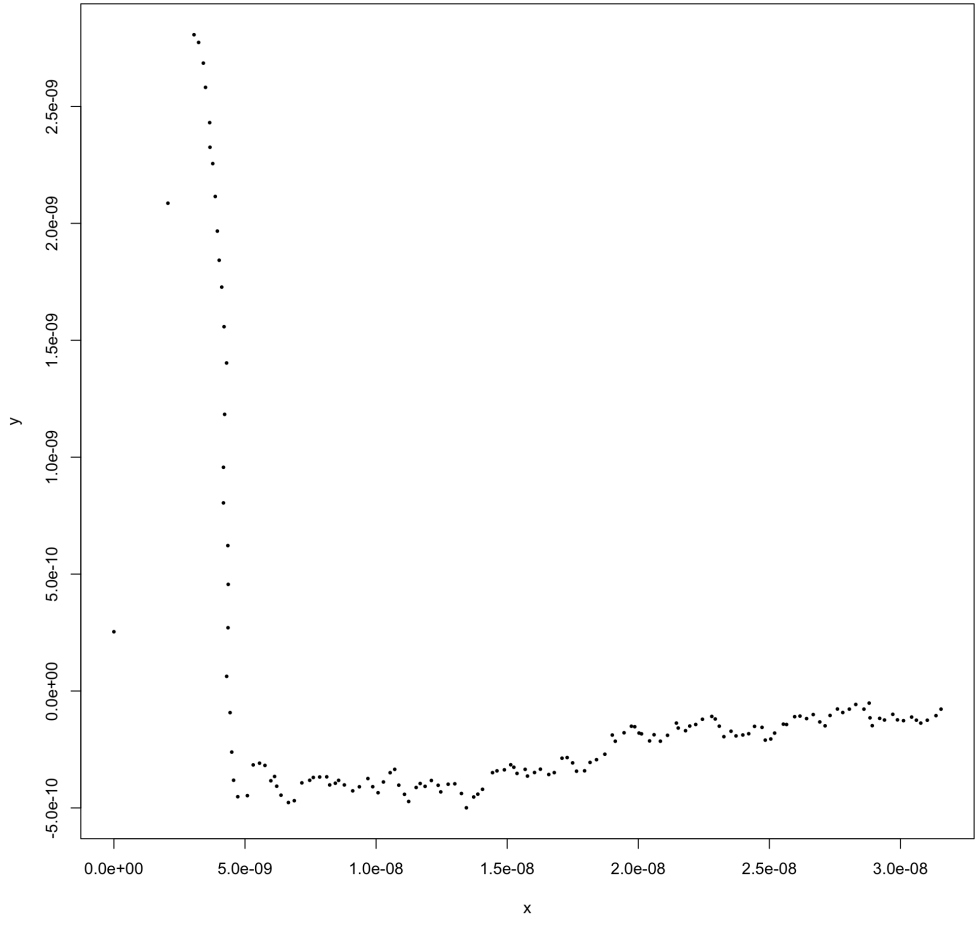
Modelling Assumptions - Aside



In this, we consider the Young's modulus as an indicator of the properties of the material – this comes with modelling assumptions! What if the material is obviously inelastic? What if it is clearly non-linear? We will come back to this point later!

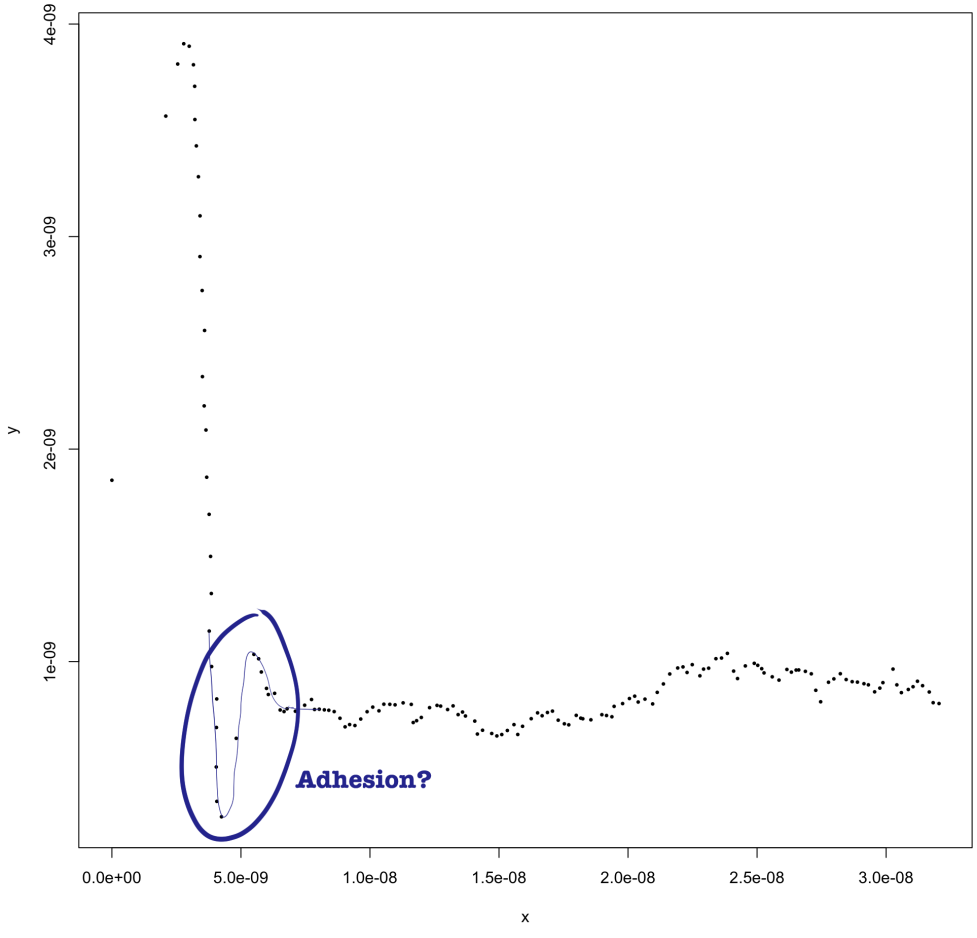
Examining Data

An ideal world has all experiment sites having the form of the previous graph.



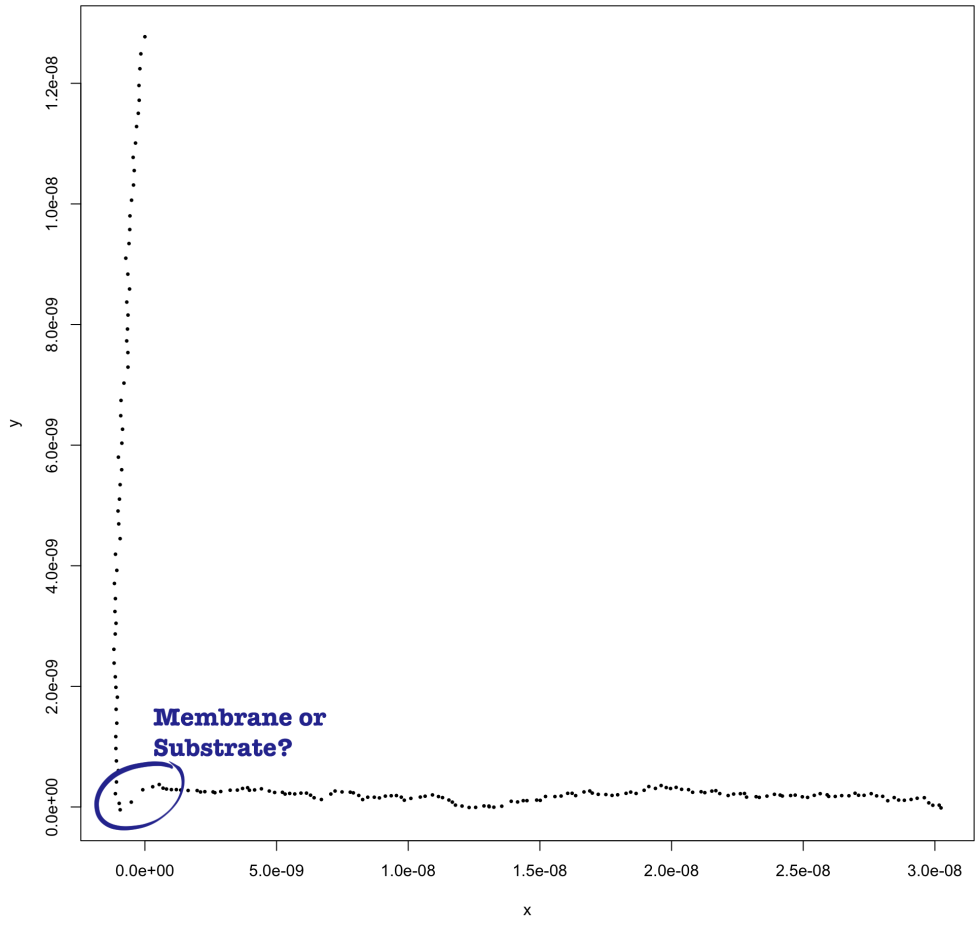
Examining Data

However, we do not live in an ideal world...



Examining Data

However, we do not live in an ideal world...



A Choice

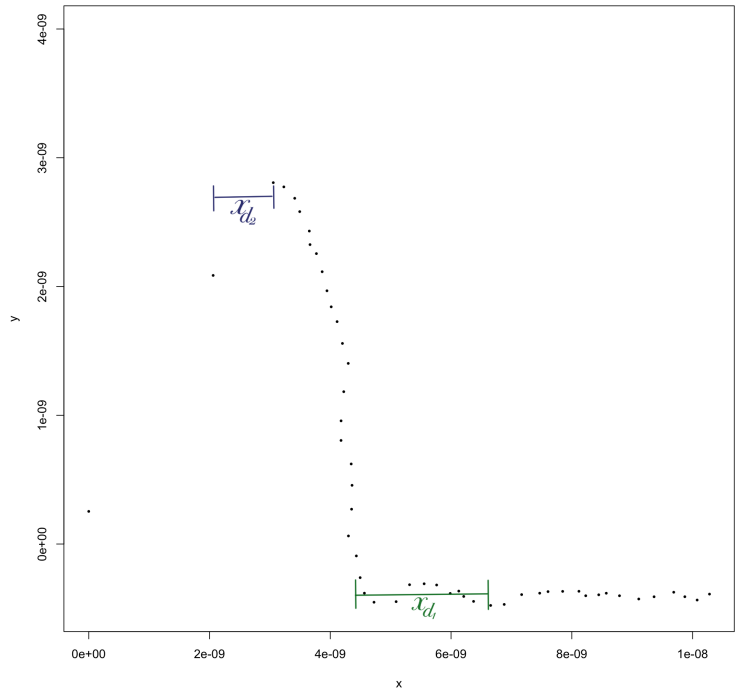
The experimental data we have is $32 \times 32 = 1024$ experimental sites. We do not want to trawl through each plot and make decisions for each individually.
I make the pragmatic decision to define:

- x_{d_1} to be the first point at which the graph turns appreciably away from $y = 0$;
- x_{d_2} to be the x-value corresponding to the maximum y-value beyond x_{d_1} and before $d = 0$.

This has the benefit of generality, and we will deal with the consequences soon.
For adhesion, maybe we could find both "peaks", and gain some insight into the adhesive behaviour across the space!

Identification Uncertainty

Even with this pragmatic choice, identification is still not straightforward.



Two different problems: noisy data makes finding d_1 difficult; sparsity of data makes the exact position of d_2 hard to find.
The difference here could be about 50% of the actual observation.

Uncertainty Quantification: d_1

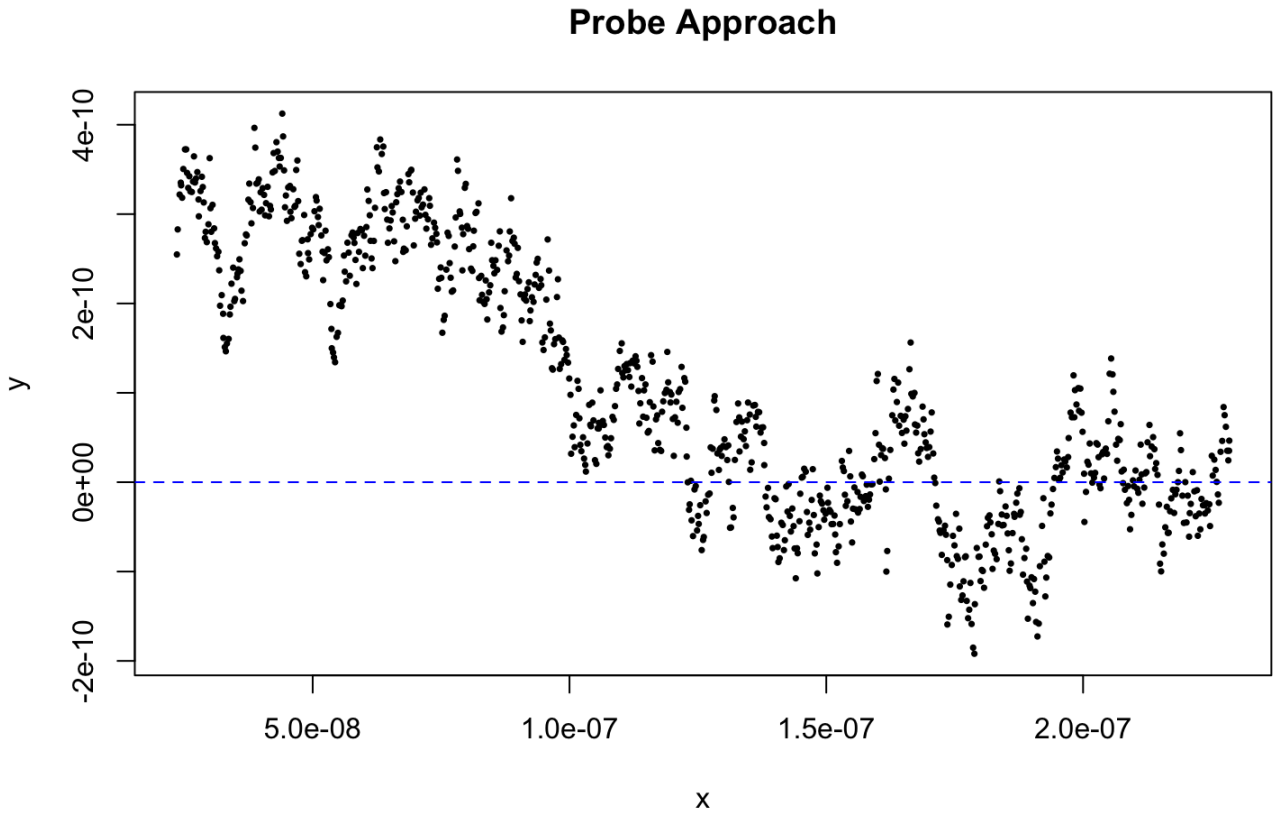
For identifying d_1, we note that the uncertainty is due to the 'wobble' in the data before the probe touches the membrane. We might make the (reasonable?) assumption that the pre-membrane data is centred around 0 with uncorrelated noise: for $d > d_1$, state that y measurements y_+ follow a Normal distribution:

$$y_+ \sim \mathcal{N}(0, \sigma^2)$$

For each experiment site, we have a large amount of data before probe-membrane connection, so the sample variance s^2 should be a good approximation of the true variance σ^2.

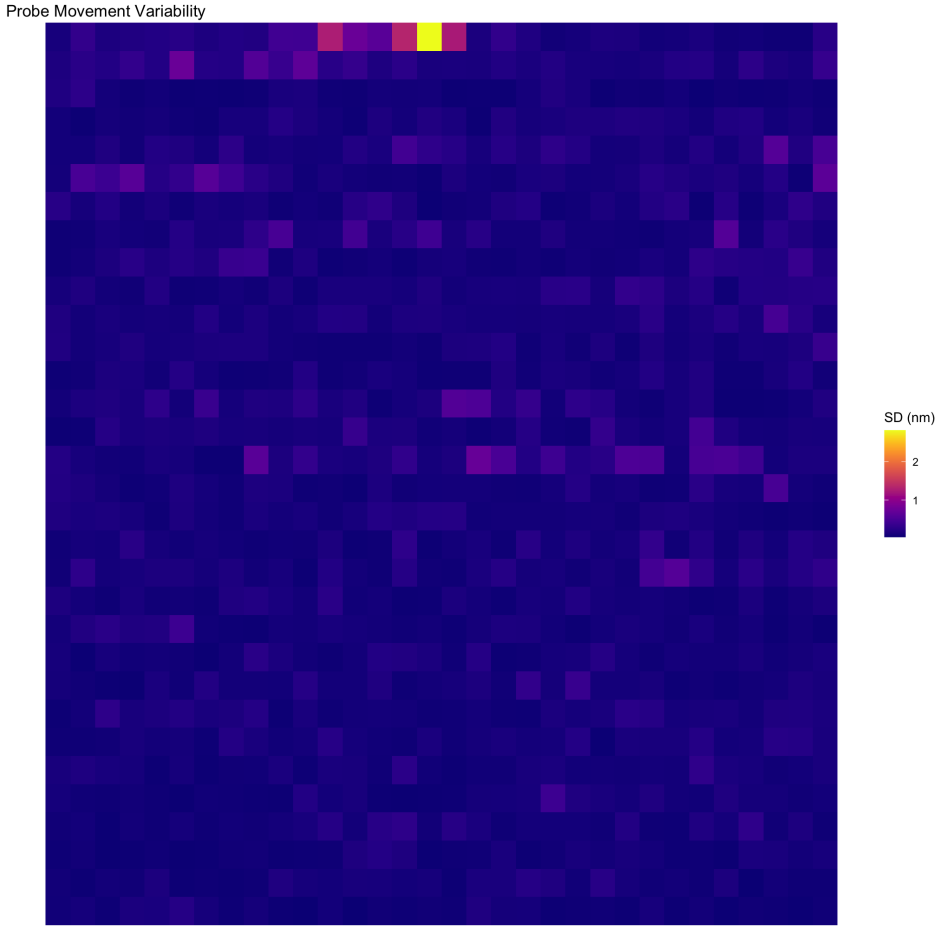
Probe Wobble

$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} y_i^2$ since $\mu = 0$. Our assumptions might not be overly convincing under investigation...



Probe Wobble

. . . but the variability due to probe approach is small and well-behaved across much of the space.



Asserting Uncertainty: d_1

The uncertainty in y due to probe approach is small, especially compared to uncertainty in distinguishing the data point to use anyway, so we can include it in our general structure using the following approach.

Finding d_1: Claim

Posit a candidate d_1 where the behaviour of the curve changes. The 'true' d_1 lies within 3 data points of the candidate point.

We encode the uncertainty in x by considering the ensemble $\{x_{d_1-3}, \ldots, x_{d_1}, \ldots, x_{d_1+3}\}$ and assuming that the range of these points corresponds to a 95% confidence interval.

Uncertainty Quantification: d_2

How do we deal with finding d_2?

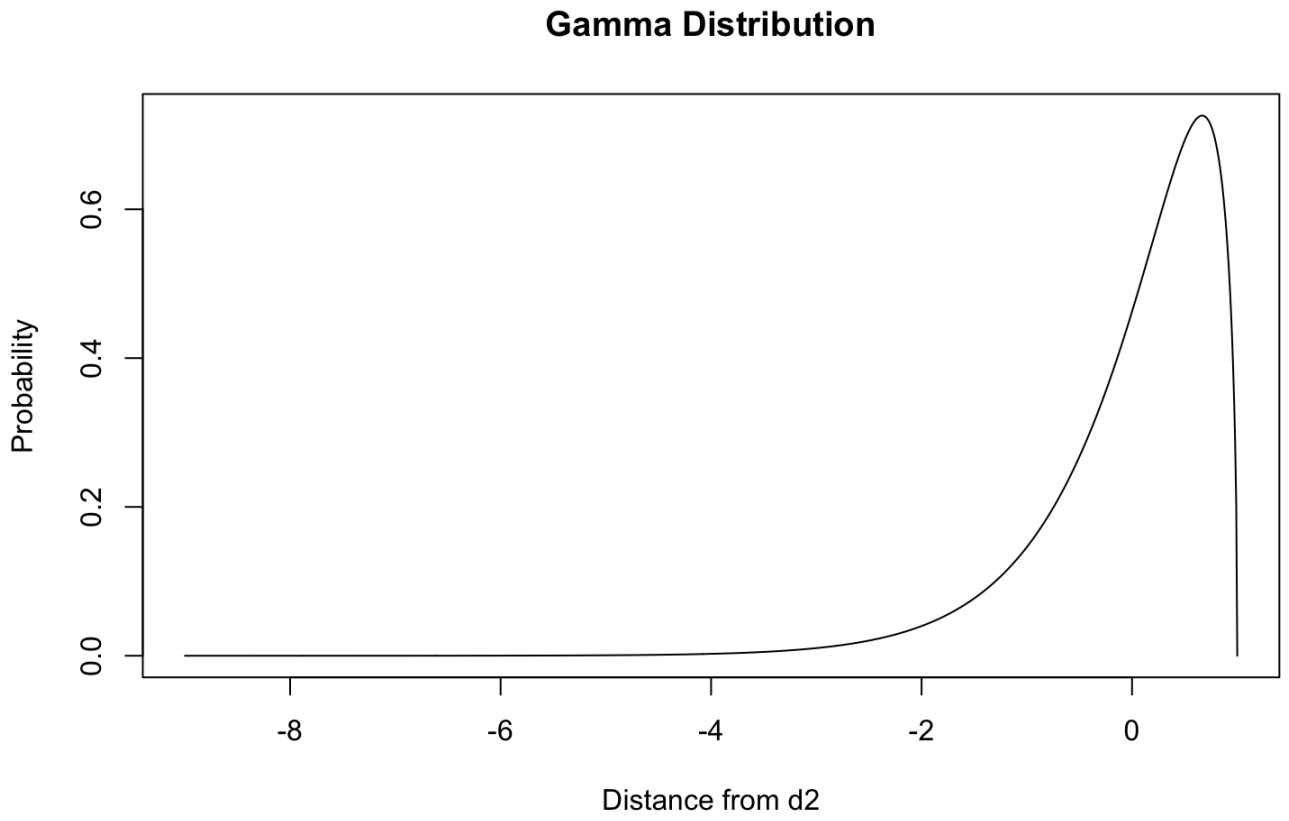
- There may be a similar 'wobble' around x_{d_2} as in x_{d_1}, but this isn't the dominant source of uncertainty.
- Our concern is whether, if we had higher resolution data, would there be a more accurate value for x_{d_2} closer to the substrate value $x = 0$.
- We only know two things: the maximal point of the peak, and the location of the next-smallest data-point x_{d_2-1}.

Assessing Uncertainty: d_2

We want to favour our determination of x_{d_2}, since our identification is (arguably) more concrete than that of x_{d_1}. We also know that any misidentification is more likely to result in us overestimating x_{d_2}, rather than underestimating it.

Assessing Uncertainty: d_2

A Gamma distribution seems appropriate: choose shape and scale parameter to get a long tail and sharp peak (here, $\alpha = \beta = 1.5$).



Obtaining Uncertainty in d_2

A Useful Result

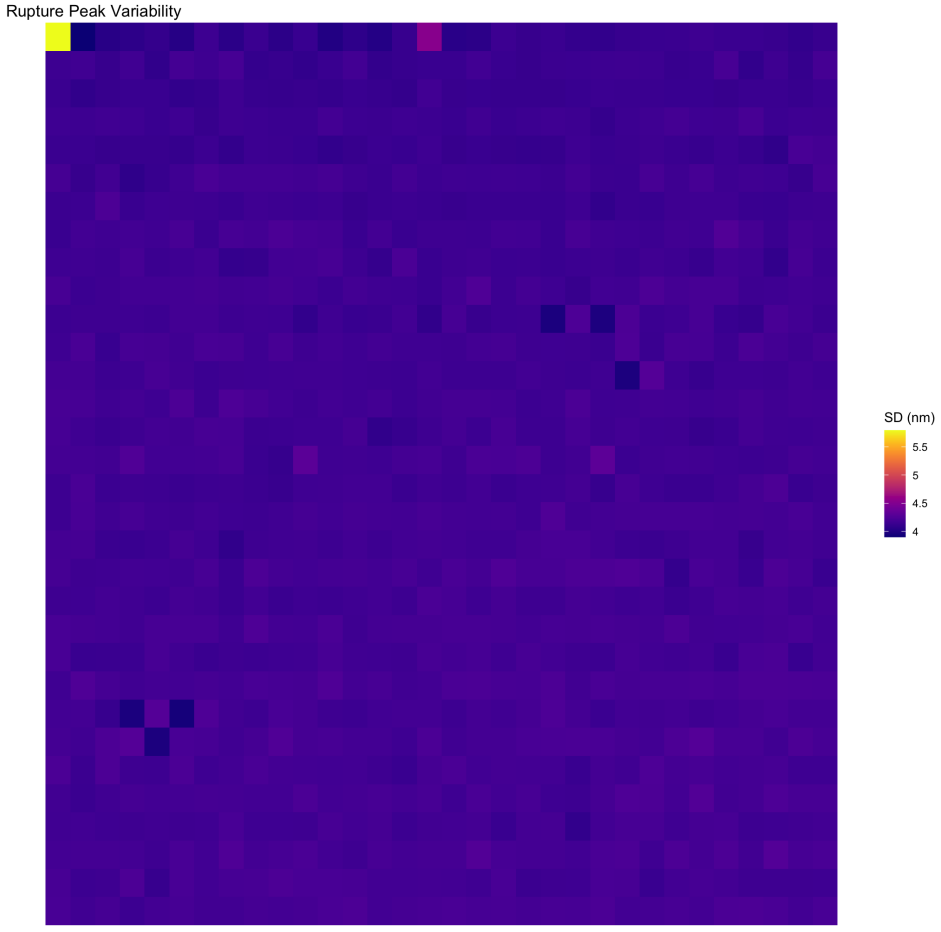
If $X \sim \text{Gamma}(\alpha, \beta)$ then $cX \sim \text{Gamma}(\alpha, \beta/c)$.

If we want next smallest observed point x_{d_2-1} to have negligible probability, then pick $X^* \sim \text{Gamma}(1.5, \beta)$ s.t. $\mathbb{P}[X = x_{d_2-1}]$ is small (say 10^{-5}).

For each experimental site, find the scaling factor c that transforms X to x_{d_2-1}, and calculate $Var[cX]$.

Rupture Peak Uncertainty

Higher uncertainty than in d_1, but uniform across the space. One anomalous value (at an edge, as with the d_1 calculation).



Induced Uncertainty in y_2

To calculate the Young modulus, we need the difference in y values, too. We need a structure for y_2, then: we *know that the Young modulus calculation assumes linearity of response*: if $Y = aX$ and $X \sim \text{Gamma}(1.5, \beta)$, then $Y_2 \sim \text{Gamma}(1.5, \beta/a)$.

We therefore calculate a using the data and our 'best guesses' for x_{d_1}, x_{d_2}, and then induce the uncertainty structure on y_2 correspondingly.

Sidenote: Another Approach

This approach uses the discrete nature of the experimental data. We could instead fit a piecewise linear model to the data, using the YM assumption:

$$f(x) = \begin{cases} 0 & \text{for } x \in [x_{d_1}, \infty) \\ a_1 x & \text{for } x \in [x_{d_1}, x_{d_2}] \\ a_2 x & \text{for } x \in [0, x_{d_2}] \end{cases}.$$

Note that we have four unknowns: a_i and x_{d_i} for $i = 1, 2$. Optimisation to find the best fit: then the uncertainty on the estimate of the Young's modulus is that of the two piecewise non-constant parts.

Young Modulus

We have D_i and Y_i for $i = 1, 2$ for every site. The Young modulus is therefore

$$E \sim \frac{Y_2 - Y_1}{X_1 - X_2}$$

All of the quantities in E are random variables. How do we estimate the uncertainty?

Simulation Study

Unfortunately, the ratio of Gamma distributions minus Normal distributions doesn't have a nice distributional expression. But we can simulate from it!

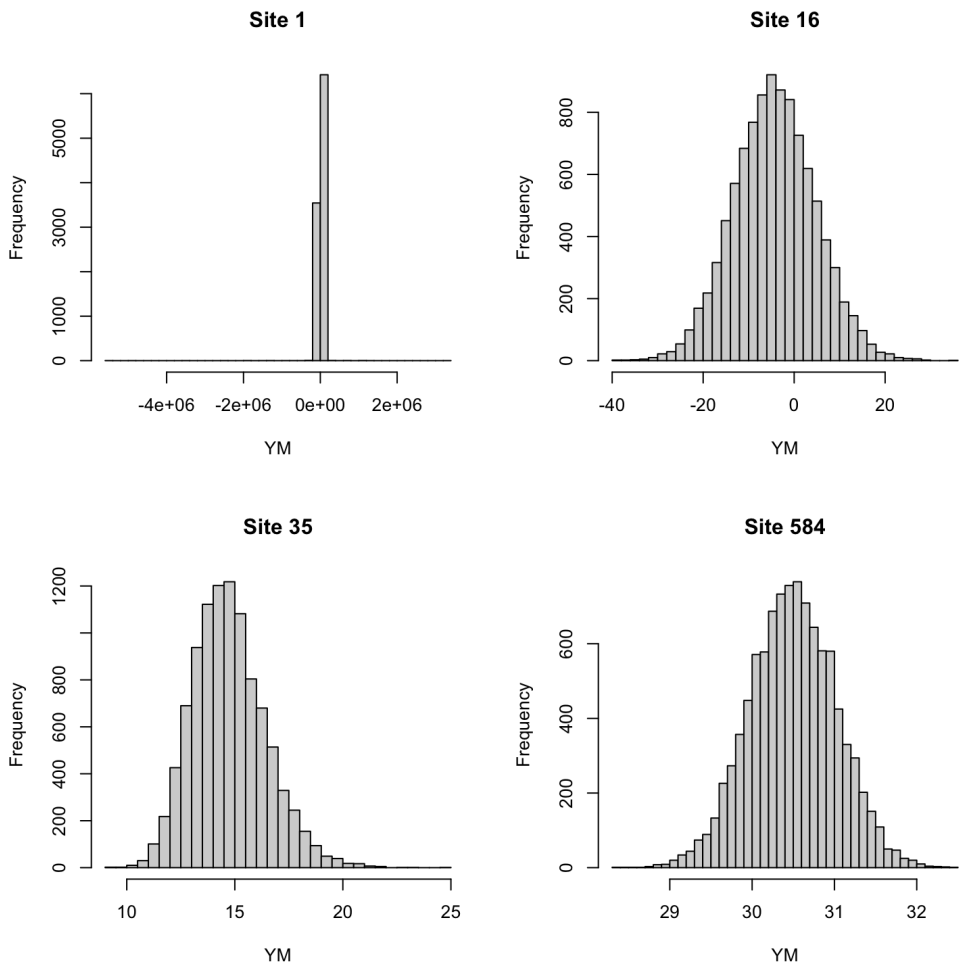
- Draw "lots" (say 10,000) samples from each distribution:

$$\{x_{1i}, x_{2i}, y_{1i}, y_{2i}\},\ i = 1, \ldots, 10000$$

- Compute the ratio $(y_{2i} - y_{1i})/(x_{1i} - x_{2i})$ for each i;
- Use these values as empirical draws from the distribution of E.

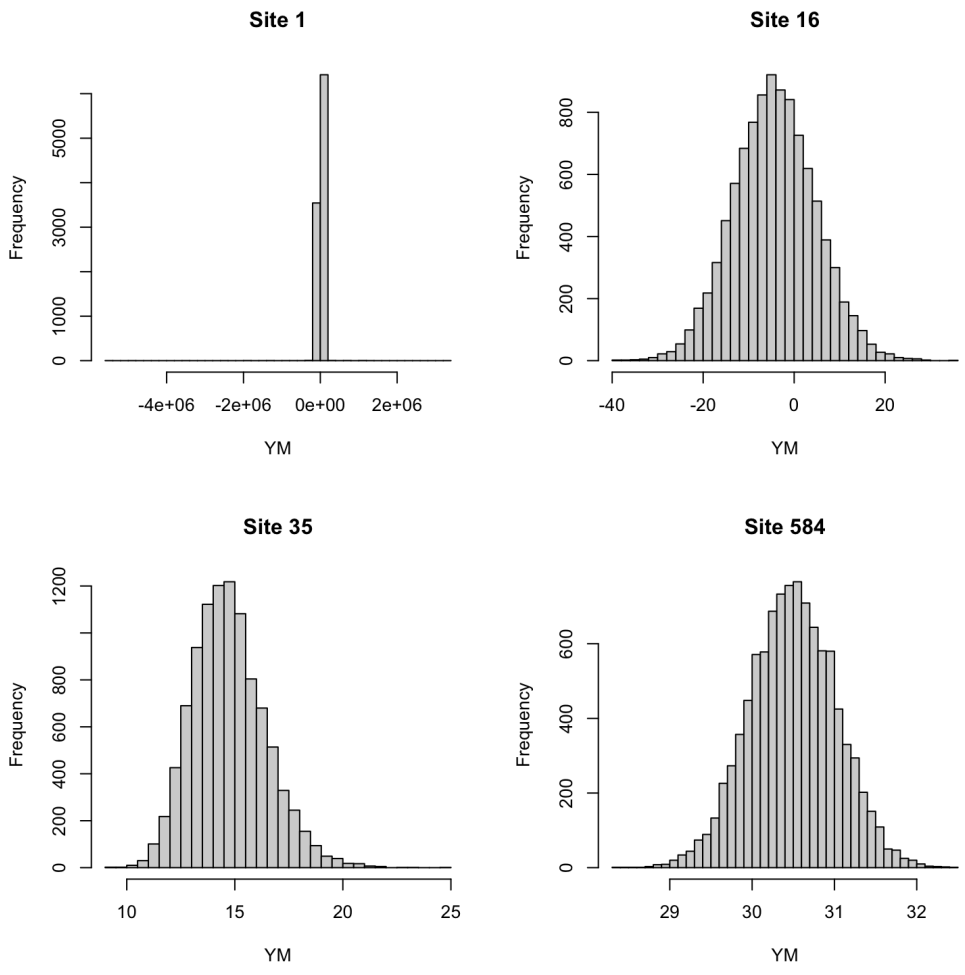
We can then find the mean, median, variance, ...

The One-Size-Fits-Most Approach



Some (in fact most) of the draws are eminently sensible.
Some have horrible outliers, and some are predominantly negative. . .

The One-Size-Fits-Most Approach



We shouldn't be too surprised. We've made a number of assumptions, and we knew they wouldn't be completely valid everywhere.
We need to examine the problem spots by-hand.

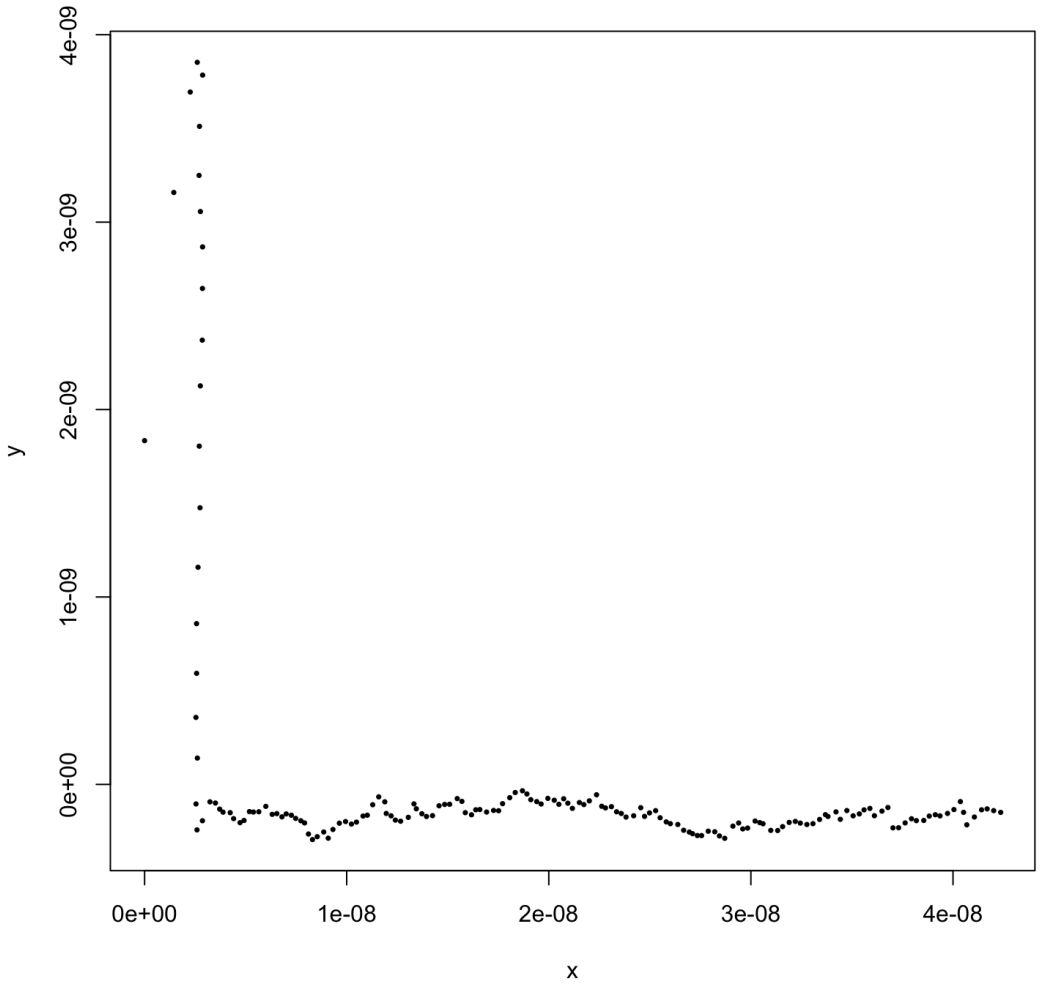
Expert Judgement

We could just discard the sites where we've obtained nonsense – but what if interesting physics is causing the breakdown in our assumptions?

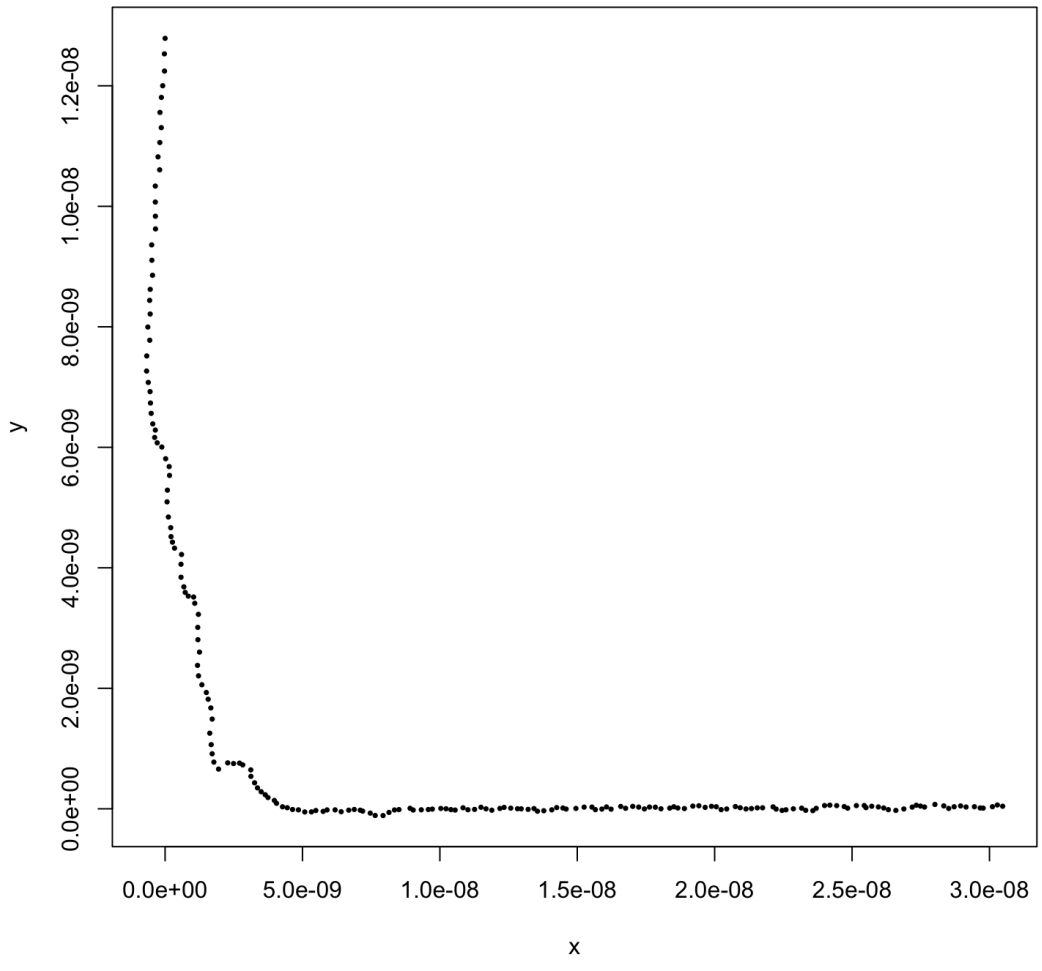
Instead, we'll go through the problem sites by-hand, identify the relevant points, put a reasonable amount of uncertainty on the determinations, and redraw E at these sites.

We might still get garbage, but at least we know we've done what we can.

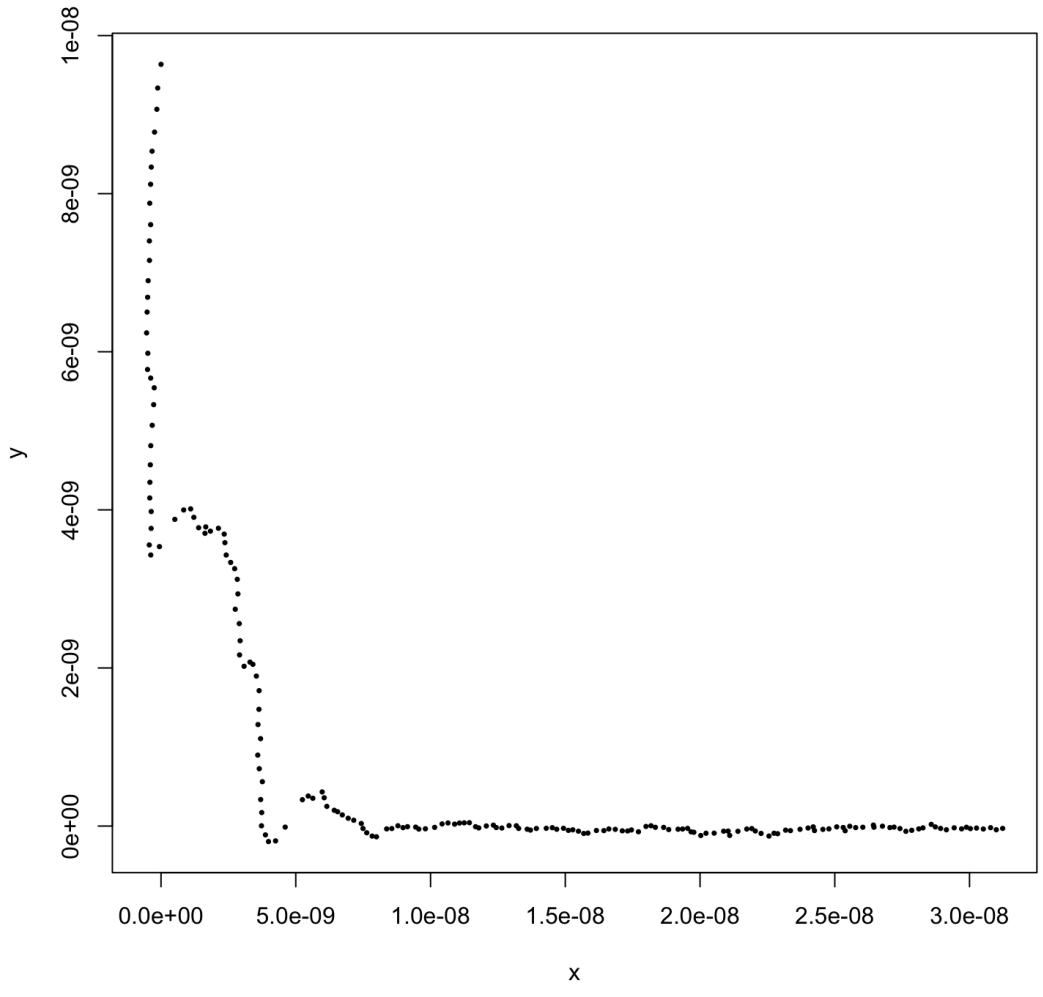
Problem Sites



Problem Sites



Problem Sites



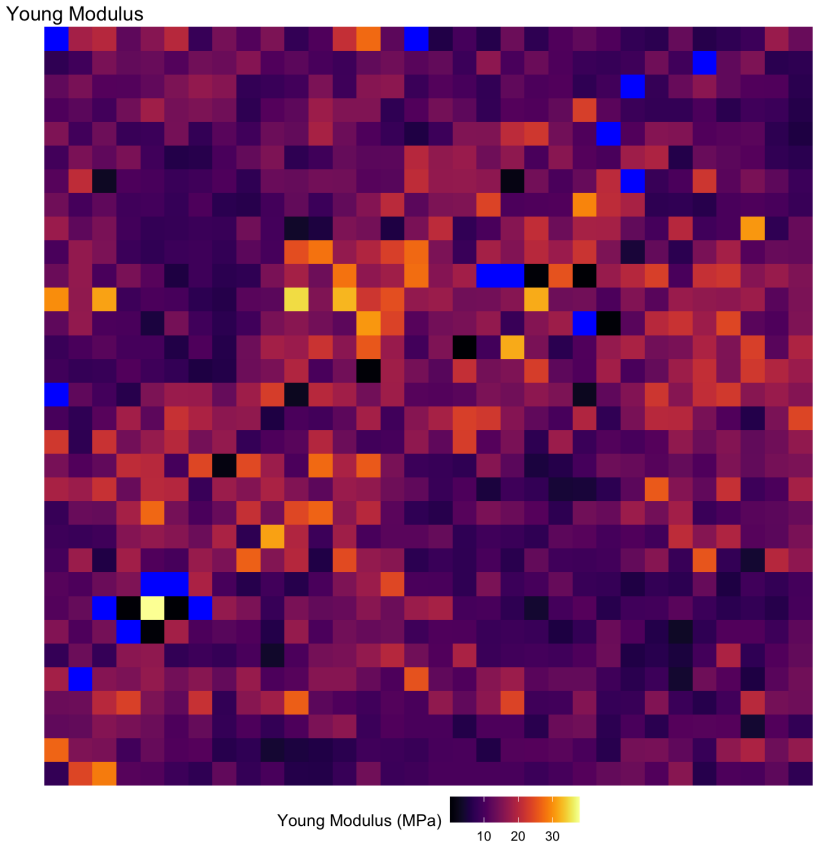
Imputation

Unsurprisingly, some of the test sites stray too far from our assumptions to allow a good determination. Despite our best efforts, the results are still not good.

If we think that the membrane should be 'well-behaved' across space, then we can try to use data imputation – we must keep note of where we've applied it, though, as the results are the least reliable we have.

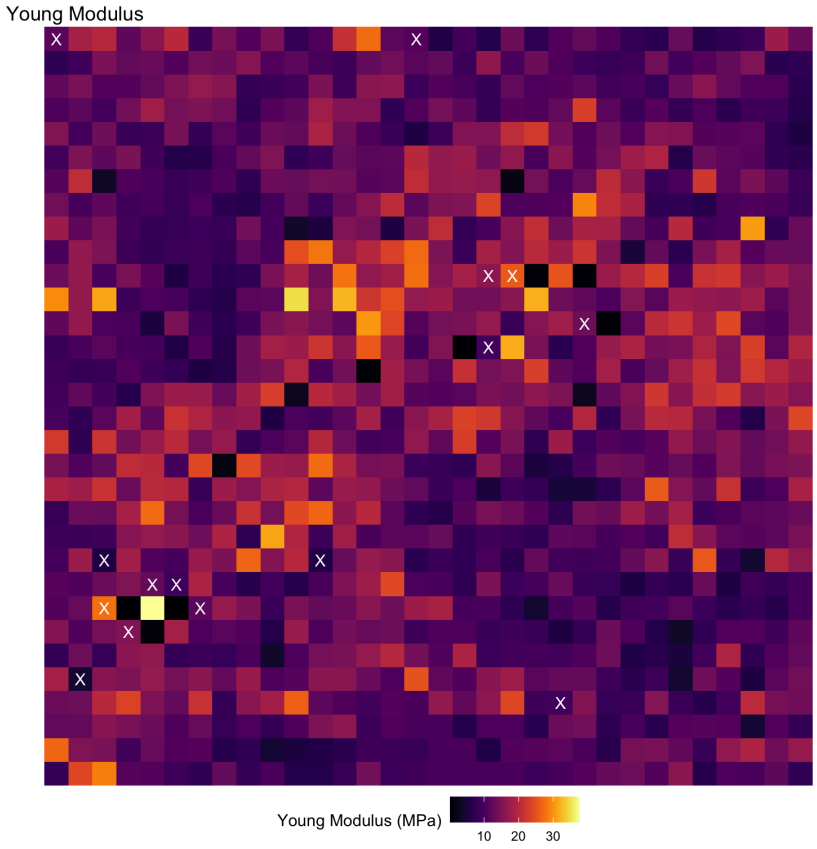
Imputation - Results

Use the nearest neighbours of a problem site to estimate the value at that point (8 neighbours in a middle cell; 5 on an edge; 3 at a corner).



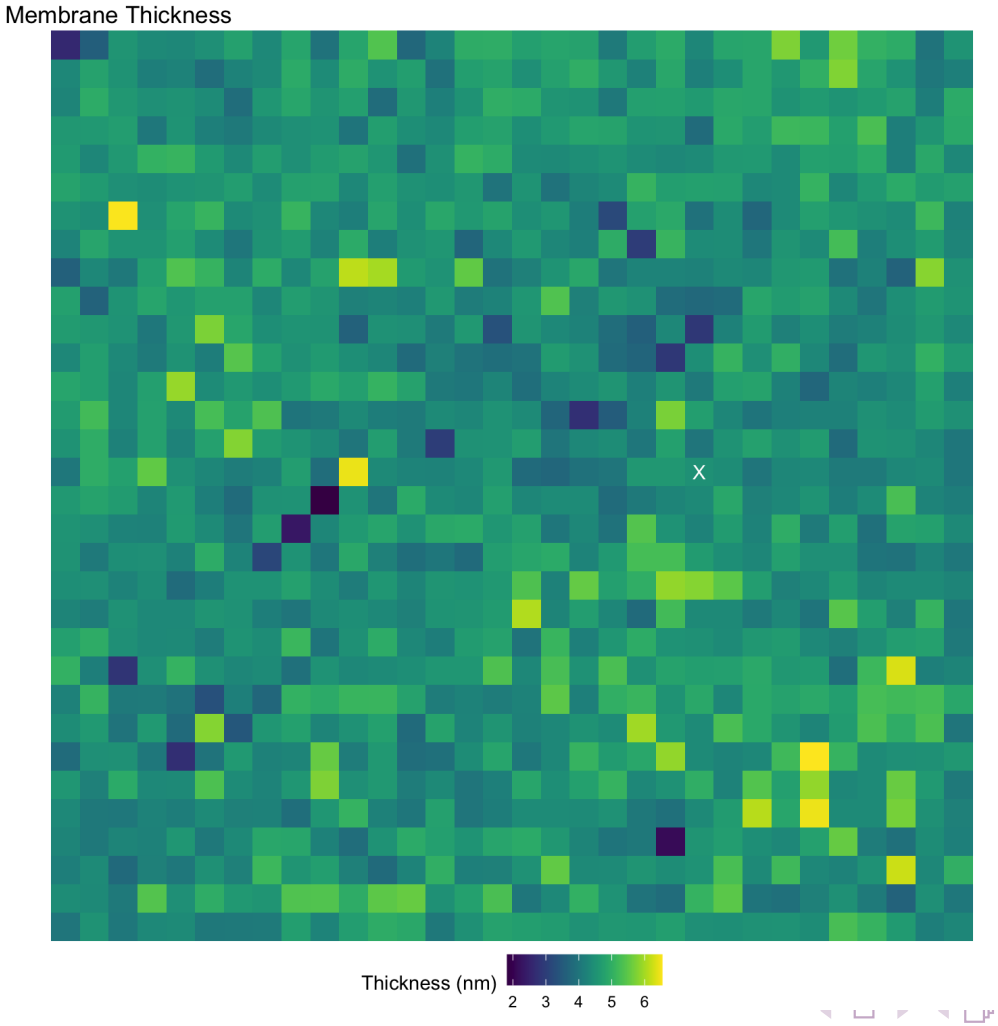
Imputation - Results

We can also weight the estimation by the original, based on how (un)reliable our original value was (maintain some structure while regularising the result).



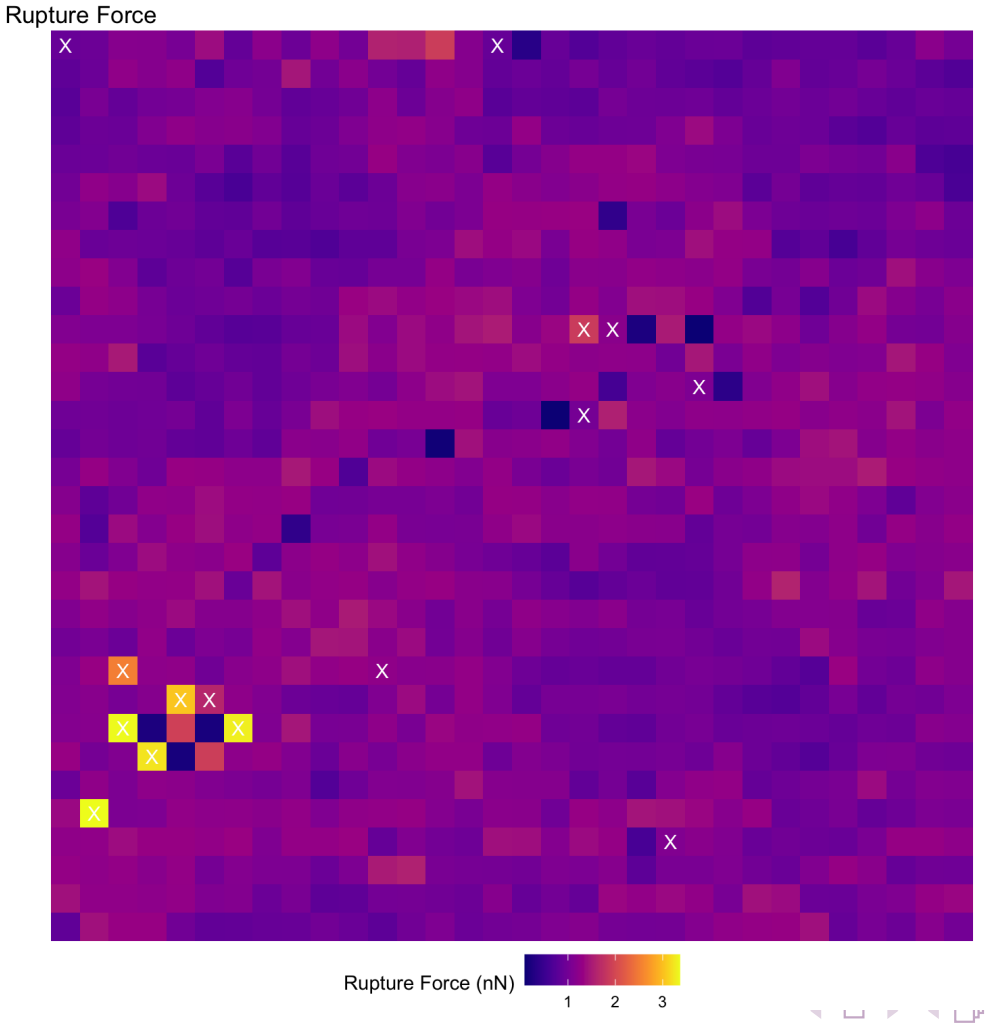
Thickness and Rupture

Play the same game with the membrane thickness and rupture force. . .



Thickness and Rupture

Play the same game with the membrane thickness and rupture force. . .



Summary – Modelling Framework

- Find the values of interest at each site;
- Quantify the uncertainty we have in these determinations;
- Propagate the uncertainties through to the quantities of interest;
- Visually inspect any anomalous results; consider violation of assumptions;
- If desired, perform data imputation.

Statstical Tests: Problem Framing

We now have clean data *and* and understanding of the uncertainties at every site.

We can now start to answer questions with some confidence:

- What parts of the membrane have 'interesting' properties?
- Are there regions where our physical assumptions break down?
- Can we say anything about the general structure were we to perform a similar experiment?

Here, we will look at substantial deviations from "background".

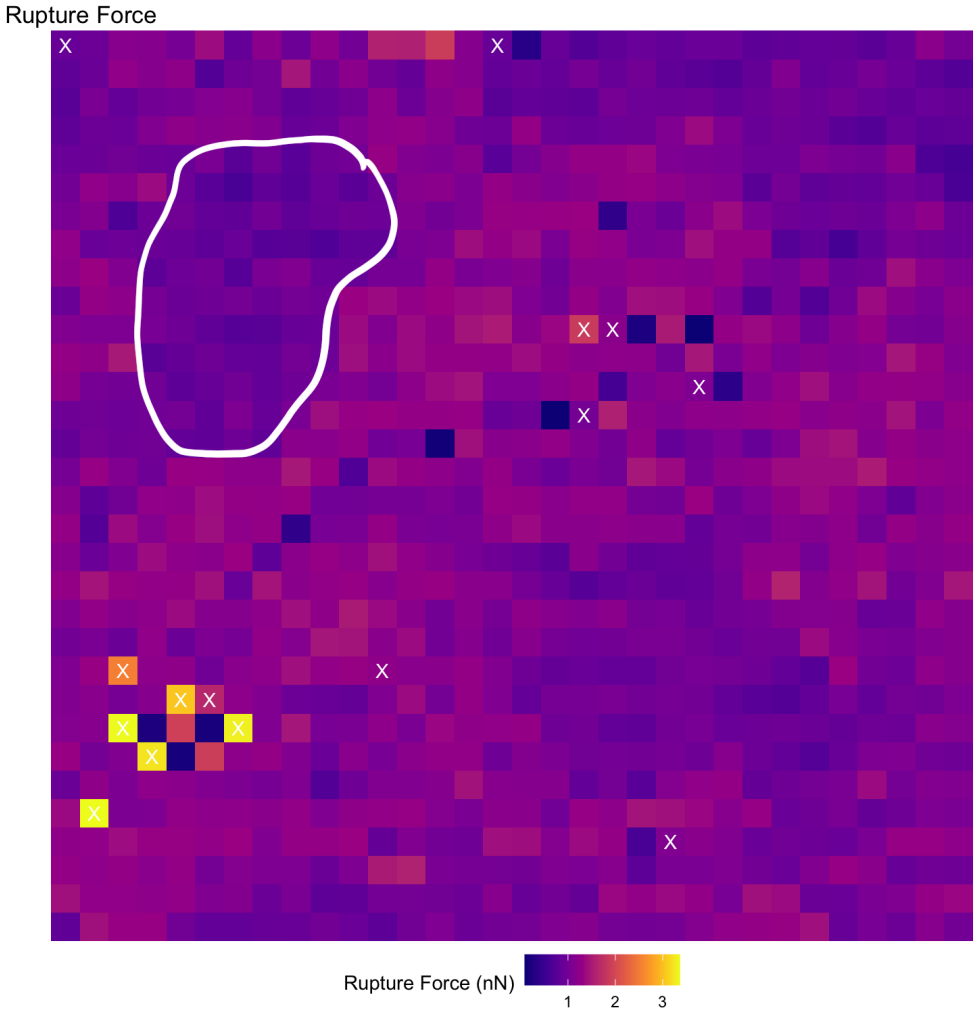
Background Definition

If we want to identify deviations from a background signal, we need to know what the background is!

Expert judgement might come in, here – perhaps we've performed control experiments to assess what the background might be. If so, we can obtain a mean estimate μ and uncertainty σ to test with.

I am not an expert, so sampled a 50-cell section of the site that looks like background to calculate the statistics.

Sampled Background



Not ideal – but using the sample standard deviation means that my smaller dataset will be accounted for. For rupture force:

$$\mu = 0.901\text{nN}$$

$$\sigma = 0.131\text{nN}$$

Testing Significance

Significance is a loaded word, but it serves the purpose. What we mean henceforth is the following:

> Is the deviation from background so large that, even accounting for all the uncertainties in the system, it is still noteworthy?

Many ways to do this: focus on three simple ideas.

- Classical Significance Testing
- Empirical Confidence Intervals
- Implausibility

Different distributional assumptions in each of these – very context dependent.

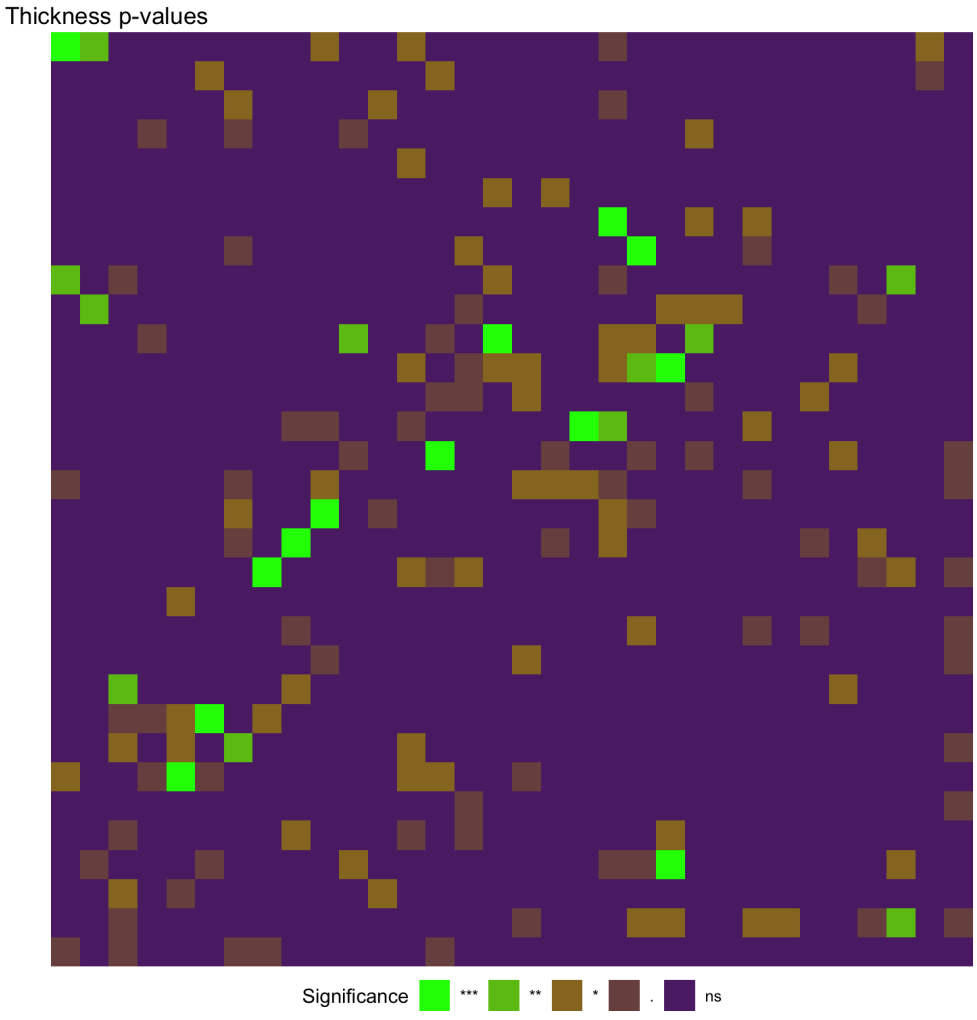
Significance Testing

We *a priori* stated that the random variable corresponding to thickness was normally distributed. If μ_B is the mean background thickness with standard deviation σ_B, and μ_i, σ_i are the mean and standard deviation of the thickness at site i, then under the hypothesis that the means are not different

$$Z = \frac{\mu_i - \mu_B}{\sqrt{\sigma_i^2 + \sigma_B^2}} \sim \mathcal{N}(0, 1).$$

These are standard Z-scores: compare to critical values of the Normal distribution to get significance values.

Significance: Thickness



Empirical CIs

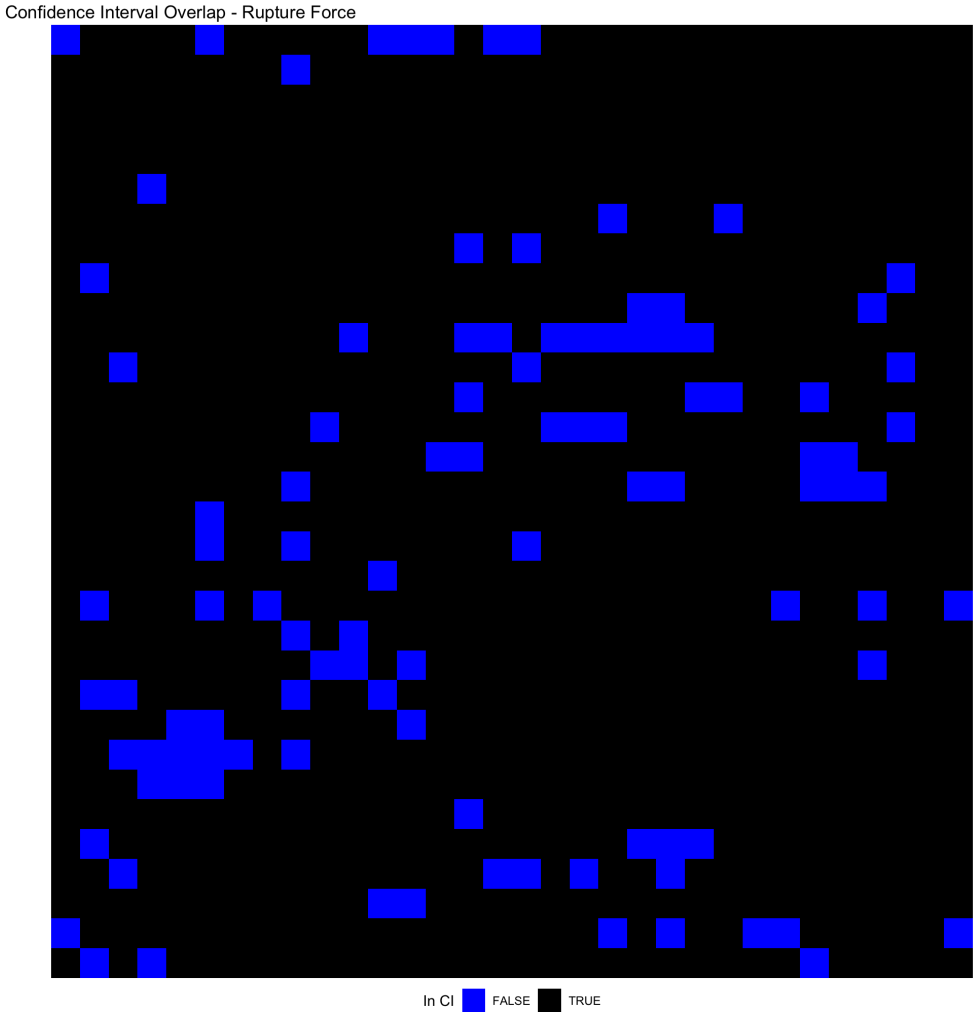
We have draws of the rupture force at each site from our sampling previously

For N samples X_i we calculate an empirical $(1-\alpha)\%$ confidence interval as

$$\text{ECI}_i = [X_{(N\times\alpha/2)}, X_{(N\times(1-\alpha/2))}]$$

where $X_{(k)}$ is the k^{th} sorted sample.
If our background interval $[\mu - n\sigma, \mu + n\sigma]$ does not overlap with ECI_i for chosen n, we consider this significant at that site.
One-way: if we only want sites where the rupture force is higher, then we check if $\mu + n\sigma \geq X_{(N\times\alpha/2)}$.

Empirical CIs: Rupture Force ($n = 3$)



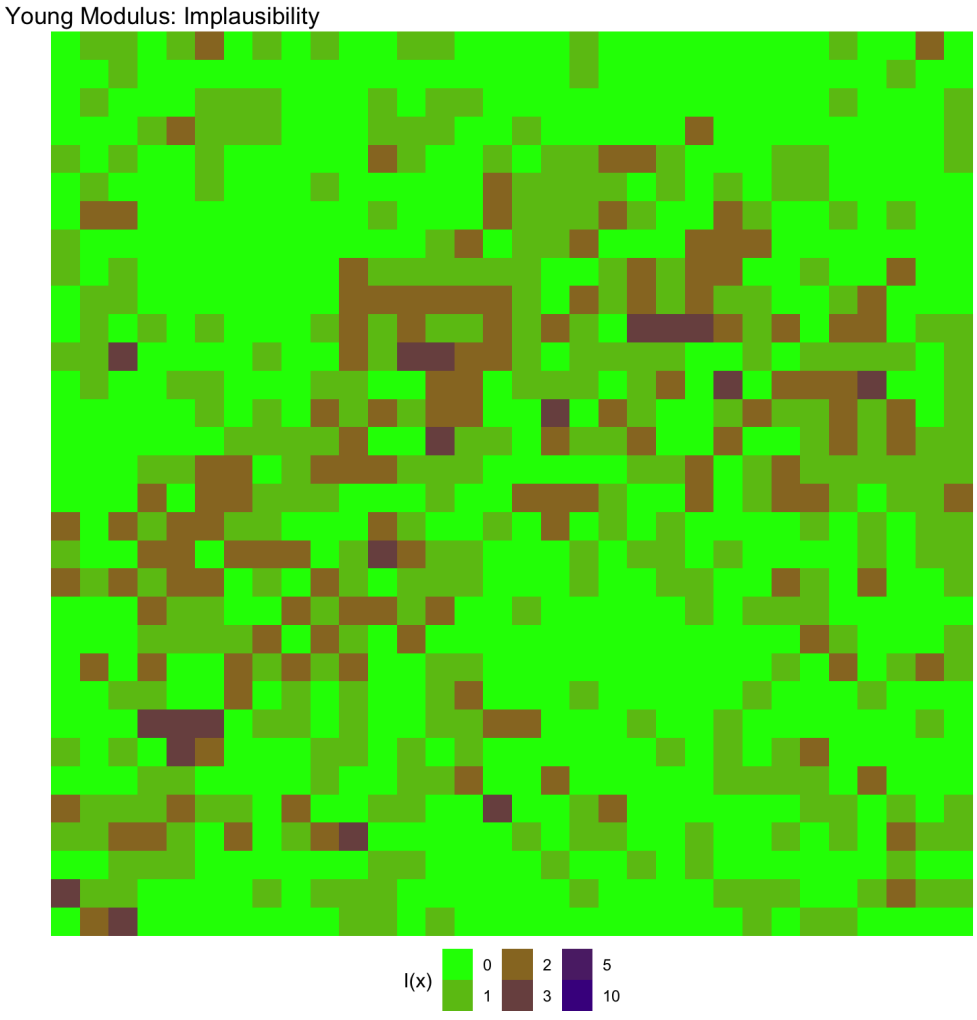
Implausibility

Implausibility is a concept drawn from *history matching* that requires no distributional assumptions: define implausibility $I(x)$ at site x for an output quantity $f(x)$ as

$$I(x)^2 = \frac{(\mathbb{E}[f(x)] - \mu)^2}{\mathrm{Var}[f(x)] + \sigma^2}$$

If $I(x)$ is "large", then there is a signficant difference between the background and our observation. If it is "small", then either the observation is close to the background *or* the uncertainty at that site is large.

Implausibility: Young Modulus



Combining Measures

Depending on the research problem, we might want to identify points where:

- The behaviour is such that imputation/manual handling was necessary;
- The thickness is significantly larger than the background;
- The rupture force is significantly higher than the background;
- The Young modulus is significantly higher than background.

Could require all/some conditions satisfied, or count how many are satisfied for each site.
We might also want the opposite: significantly thinner/weaker membranes. If so, we simply reverse some of the arguments above.

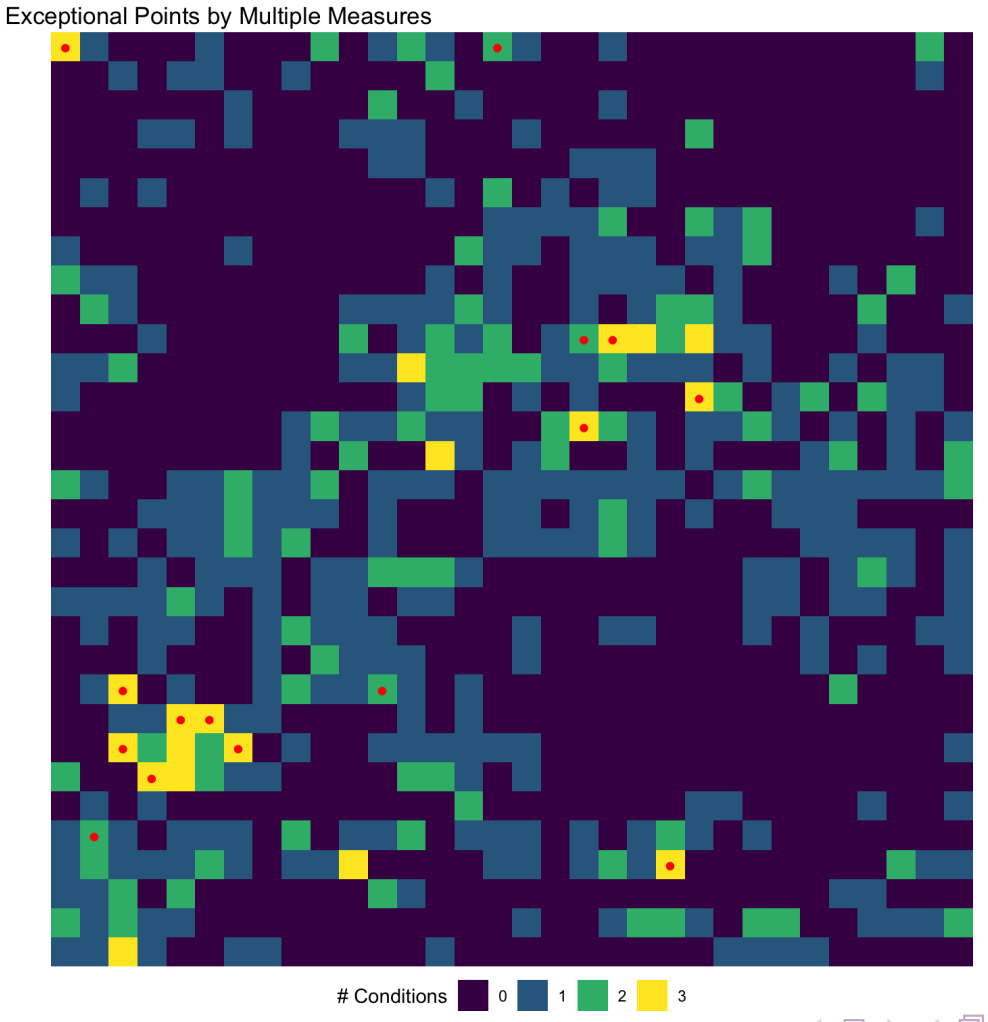
Set Definitions

Define sets

- O: Points at which manual identification or imputation was necessary;
- P_α: Points with thickness p-value smaller than α;
- CI_n: Points with non-overlapping force confidence intervals with background $[\mu - n\sigma, \mu + n\sigma]$;
- I_c: Points with Young modulus implausibility greater than c.

Combined Measures – Example

$$\alpha = 0.05,\ n = 3,\ c = 3$$



UQ for SPM

- Considering data variability is crucical for ensuring results are **robust**, **reproducible**, and **reliable**.
- Expert judgement/careful consideration of visualisations allows us to make best use of summary statistics **and** reasonably quantify our uncertainties.
- Once we understand the structure of the data we have, any uncertainties can be carried through to statistical testing, both empirically and parametrically.

Statistical Relevance

Points that have statistically significant differences as a result of our tests can be considered and can aid further experimental design:

- Boundary effects – are edge locations more likely to display anomalies? Can we design experiments to combat this, if so?
- Are impurities in the medium causing significance? Extreme values might be due to properties of the medium or a probe failure at a given site.
- Multiple experiments: if we assume a common background medium, then what differences can we determine from two experiments? Tests for homogeneity (χ^2, KS, Wilcoxon,...)
- Identification: if we have statistically appropriate conditions under which a site is classed as 'anomalous', we can create a classification scheme for future experiments.

Take-Home

Statistical tools can give us ways to justify intuition or expert judgement formally. Different tools to deal with parametric (i.e. we know the distribution) and non-parametric (no known/tractable distribution) data, but we can always quantify the data.

They are not a complete replacement for expert judgement! Inspect your data, consider your plots, and use your knowledge to guide the determinations you make (all the easier to defend them).

Finally. . .

1. Plot (a subset of) your data – persuade yourself the summary statistics you plan to use are meaningful for the problem at hand.

Finally...

1. Plot (a subset of) your data – persuade yourself the summary statistics you plan to use are meaningful for the problem at hand.
2. Consider the process used to extract the summaries: identify uncertainties in using them.

Finally...

1. Plot (a subset of) your data – persuade yourself the summary statistics you plan to use are meaningful for the problem at hand.
2. Consider the process used to extract the summaries: identify uncertainties in using them.
3. Attach distributions to the quantities or, where not possible, use a simulation study to empirically evaluate.

Finally...

1. Plot (a subset of) your data – persuade yourself the summary statistics you plan to use are meaningful for the problem at hand.
2. Consider the process used to extract the summaries: identify uncertainties in using them.
3. Attach distributions to the quantities or, where not possible, use a simulation study to empirically evaluate.
4. Examine outliers/anomalies – impute if it is statistically justified.

Finally...

1. Plot (a subset of) your data – persuade yourself the summary statistics you plan to use are meaningful for the problem at hand.
2. Consider the process used to extract the summaries: identify uncertainties in using them.
3. Attach distributions to the quantities or, where not possible, use a simulation study to empirically evaluate.
4. Examine outliers/anomalies – impute if it is statistically justified.
5. Perform statistical testing for significance using the summary quantities and the associated uncertainties.

Finally...

1. Plot (a subset of) your data – persuade yourself the summary statistics you plan to use are meaningful for the problem at hand.
2. Consider the process used to extract the summaries: identify uncertainties in using them.
3. Attach distributions to the quantities or, where not possible, use a simulation study to empirically evaluate.
4. Examine outliers/anomalies – impute if it is statistically justified.
5. Perform statistical testing for significance using the summary quantities and the associated uncertainties.
6. Break any of these rules sooner than do anything barbarous.

Another Approach - Emulation

An *emulator* is a statistical surrogate for a physical process or model. It predicts at unseen points in parameter space and encodes its uncertainties about those predictions.

The general structure is

$$f(x) = \text{Global Regression Term} + \text{Local Variation}$$

Another Approach - Emulation

An *emulator* is a statistical surrogate for a physical process or model. It predicts at unseen points in parameter space and encodes its uncertainties about those predictions.

The general structure is

$$f(x) = \sum_{j=1}^{p} \beta_j h_j(x) + u(x)$$

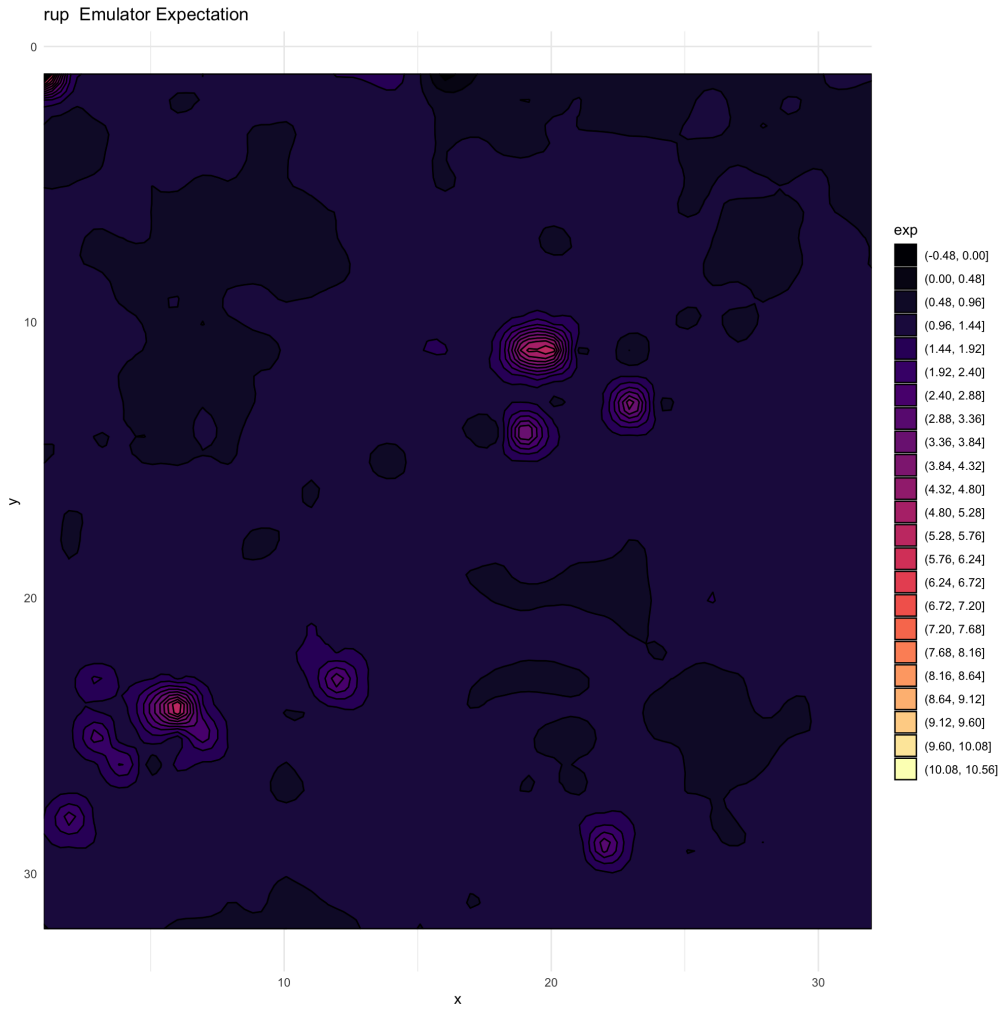
Emulation - Update

Set up prior specifications: namely, expectation $\mathbb{E}[f(x)]$ and uncertainty $\mathrm{Var}[f(x)]$, as well as correlation between different points $\mathrm{Cov}[f(x), f(x')]$. Then when given data D, obtain posterior predictions

$$\mathbb{E}_D[f(x)] = \mathbb{E}[f(x)] + \mathrm{Cov}[f(x), D]\mathrm{Var}[D]^{-1}(D - \mathbb{E}[D]),$$
$$\mathrm{Var}_D[f(x)] = \mathrm{Var}[f(x)] - \mathrm{Cov}[f(x), D]\mathrm{Var}[D]^{-1}\mathrm{Cov}[D, f(x)]$$

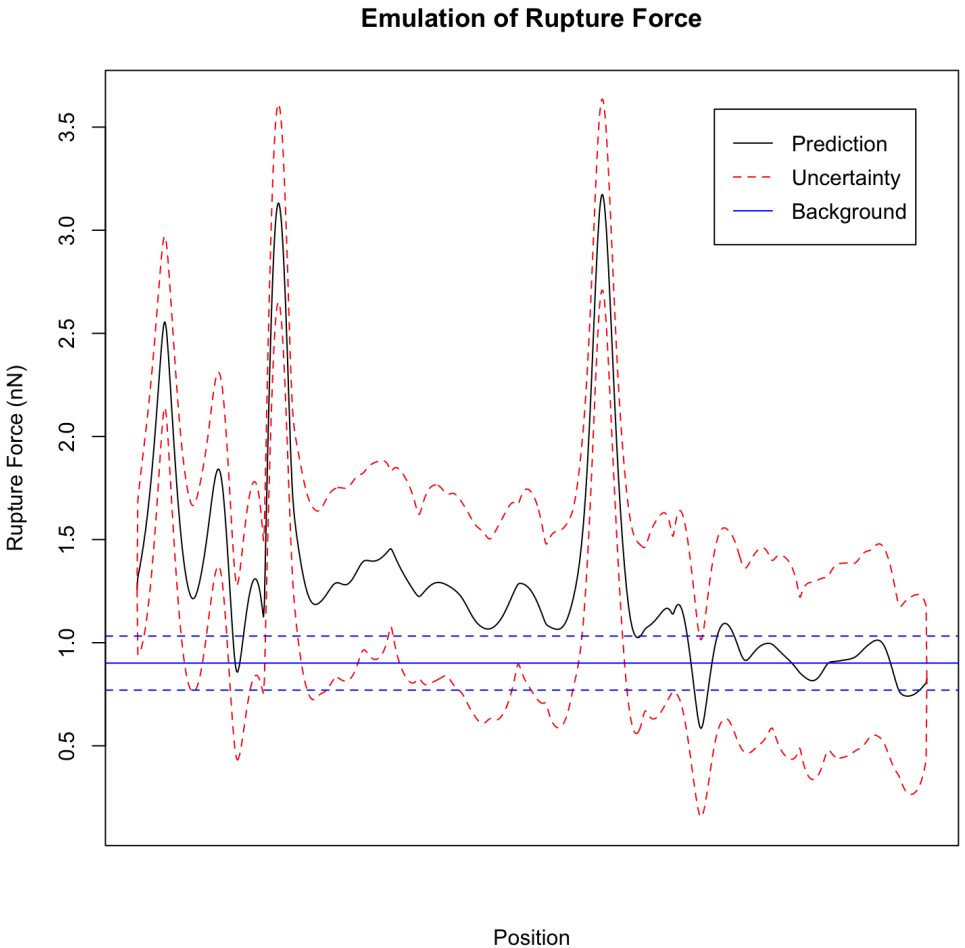
Using the *Bayes linear framework here: no explicit distributions required for the posterior update.*

Emulation - Prediction



Emulation - Implausibility

Can use a measure like implausibility to assess substantial deviations from baseline.



One-dimensional projection along a 'line of interest': running from bottom-left $(1, 29)$ to top-right $(32, 4)$. 'Significant' points have no overlap between the blue and red errorbars.

Emulation - Summary

- A form of statistical interpolation, except with robust extrapolation ability.
- Prediction uncertainty comes automatically along with the predictions themselves.
- No need to specify distributions; distribution-free measures can be applied to predictions to determine significant deviations.
- Similarly dependent on prior specifications – expert judgement still matters!
- More complex to set up than some other methods.