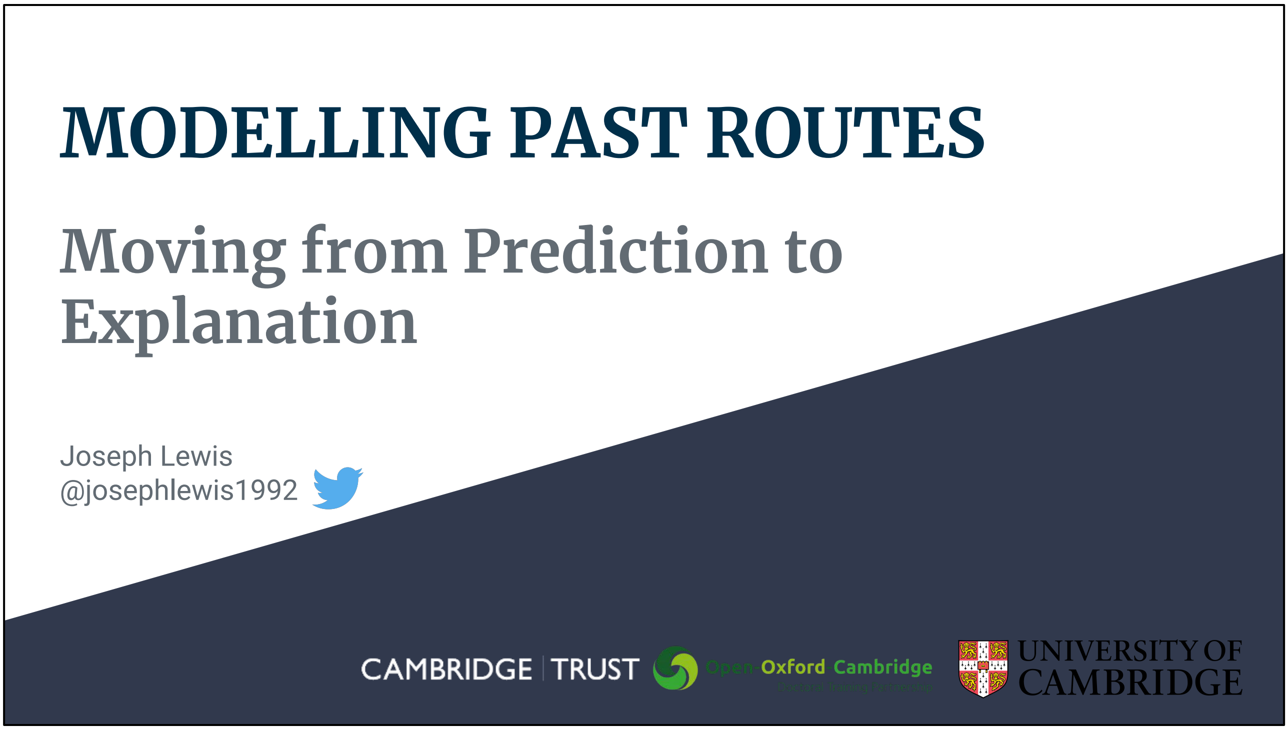


MODELLING PAST ROUTES

Moving from Prediction to Explanation

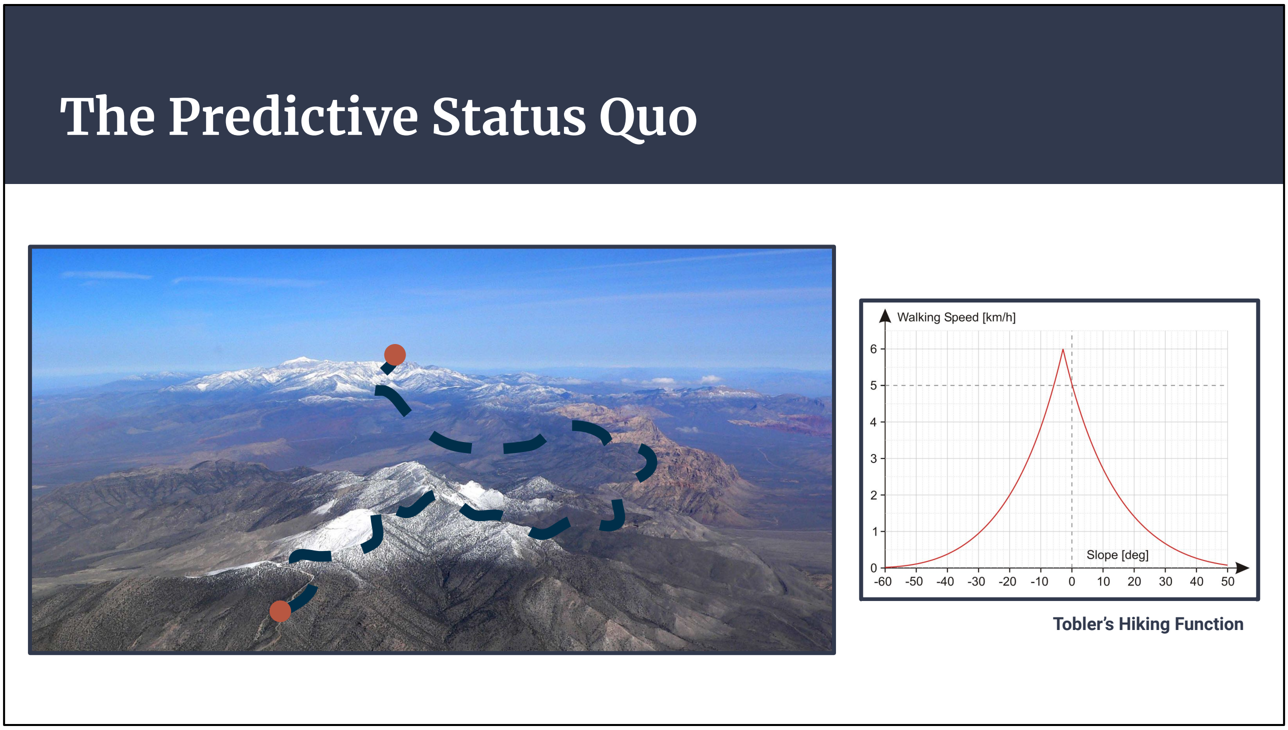
Joseph Lewis
@josephlewis1992

CAMBRIDGE | TRUST
Open-Oxford-Cambridge
Doctoral Training Partnership
UNIVERSITY OF CAMBRIDGE

Firstly, thank you to the Digital Archaeology & Heritage Lab for the introduction and for the invitation to present my PhD research.

For my talk, I'll be presenting a new approach for how we can move from predicting where routes might have been in the past to explaining past routes. Through this, we can start to understand the decision-making processes used by past people when traversing the landscape.

I'll be using Roman roads in Britain as the case study but the approach is applicable to all routes



The Predictive Status Quo

Tobler's Hiking Function

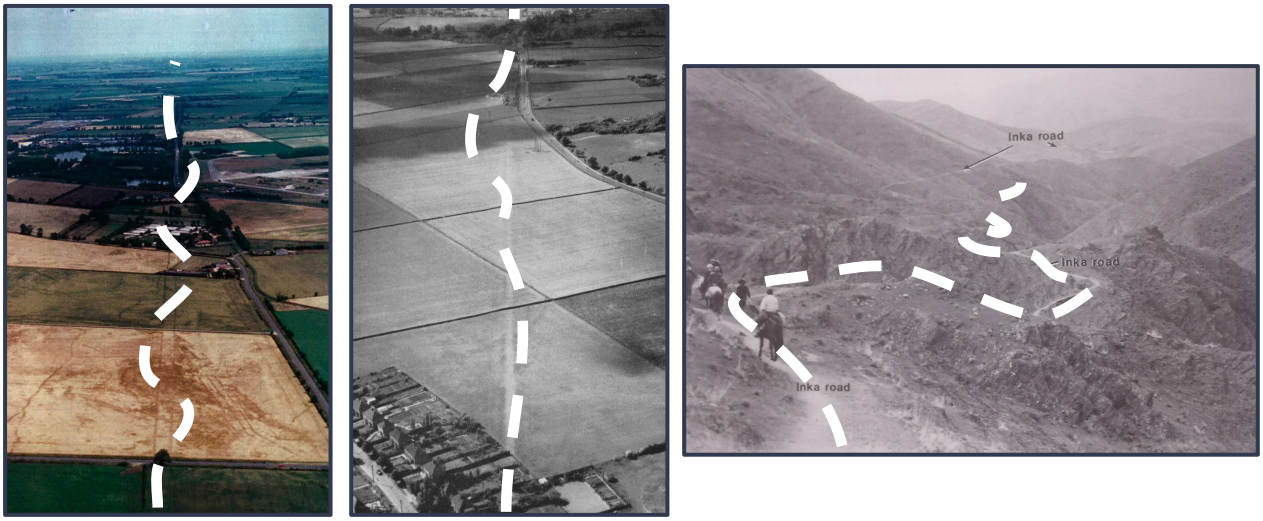
Before moving to explaining past routes, I thought it would be good to outline what I believe to be the status quo when predicting where past routes might have been.

This is normally done when we want to understand where people in the past might have moved, but we lack direct record for where they moved.

Within this predictive approach, we have two locations that we want to connect. Under the assumption that humans minimise some cost when traversing the landscape, with cost often measured in terms of time or energy, for example Tobler's hiking function on the right, we calculate the least costly path from the origin location to the destination location.

This least cost path then represents the path that people might have taken when traversing between these two locations

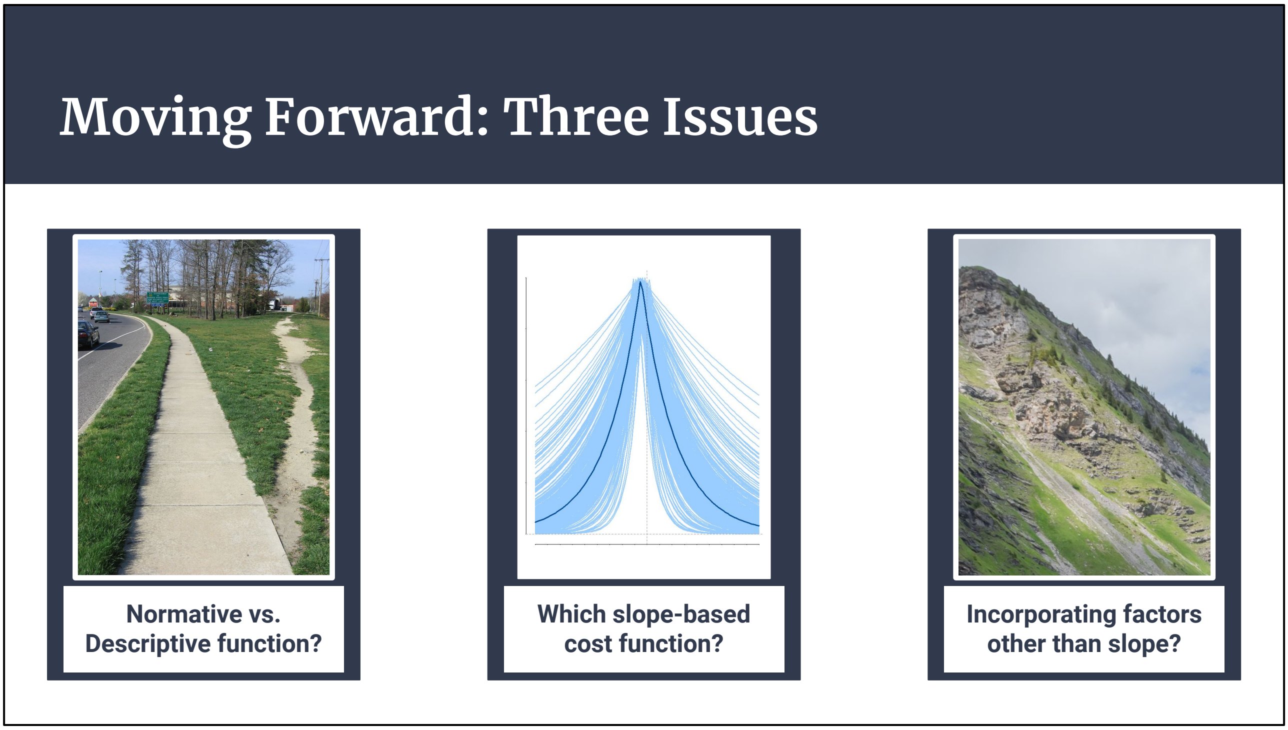
Moving from Prediction to Explanation



When moving from predicting where people might have moved to explaining known past routes, it’s common for the same approach to be applied: that is, calculating a least-cost path assuming that humans minimise cost when traversing the landscape. The added step is comparing this least-cost path, for example as shown by the white lines, to the known route. If deviations occur between the two, we can suggest that the known route did not follow the expected path under the assumption that humans minimise some cost.

And, whilst this provides a heuristic tool to say that known routes are not following expected behaviour given our assumption, this approach has multiple limitations: For example, this assumes that past people only sought to minimise a single cost when traversing the landscape, as well as providing little guidance for how we can improve the model

If you’ve done any point pattern analysis, I view this as akin to testing whether points are random or clustered. Sure, it’s good to test whether points are randomly distributed, but we known that this is unlikely to be true before doing the test. And when we’ve done the test and shown that the points are not random, we’re left in a position of no greater clarity for understanding the **Why** questions: why are the points distributed as they are, or in the case of past routes why did it take this path?



Moving Forward: Three Issues

Normative vs. Descriptive function?

Which slope-based cost function?

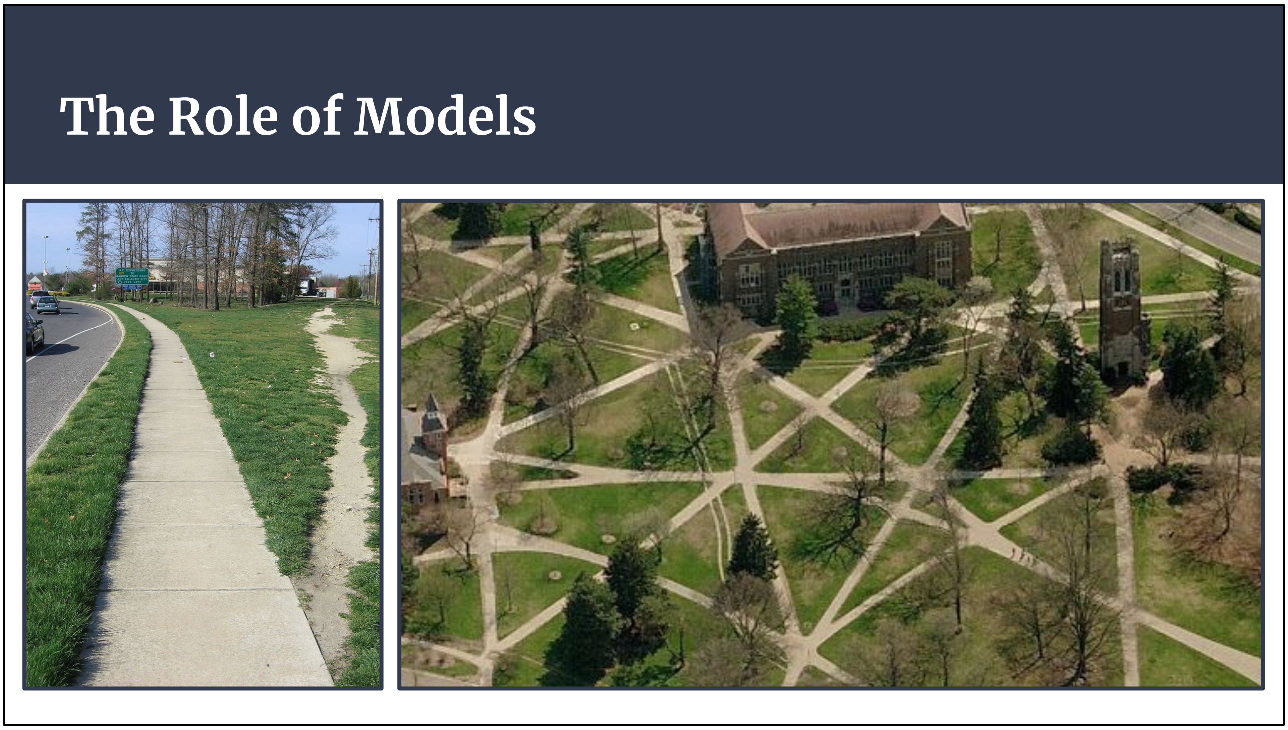
Incorporating factors other than slope?

Moving forward, I propose that three issues need to be addressed when aiming to explain known past routes

The first is what function does the model play? Is it normative whereby we model what people ought to do based on some general law, or descriptive where we aim to model what people actually did?

The second is the issue with choosing a cost function when modelling routes. For example, which cost function do we choose? And should we even aim to choose a single cost function?

And lastly, how do we incorporate additional factors other than slope into our models?

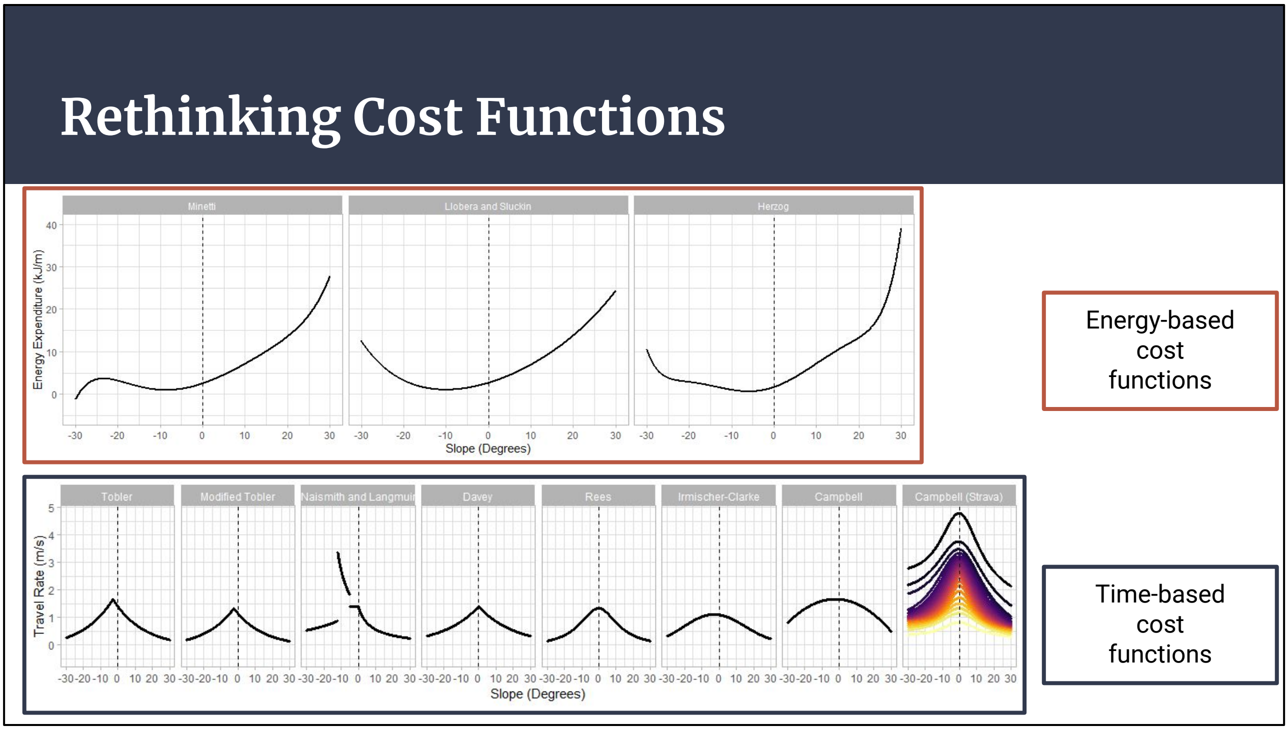


The Role of Models

First is what role should the model play when aiming to explain known past routes? As an analogy, here's a university campus. Within this campus, there are paths cutting across the field. These paths were however not laid out based on where the university wanted people to move across the field, but rather reflecting where people actually moved.

And this is how we need to think about our model when aiming to explain known past routes. We are not interested in our model playing a normative function where we aim to understand where past people ought to have moved given general laws, but rather our model playing a descriptive function, where we aim to understand where past people actually moved.

Therefore the models we create should not be based on general laws such as solely minimising time or energy, and instead be descriptive, culturally dependent, and reflect the historical, political, and social dynamics in which the routes were created in.



Rethinking Cost Functions

Energy-based cost functions

Time-based cost functions

Second issue is how we use cost functions when aiming to explain known past routes

Here we have a range of cost functions that numerically express the relationship between slope and energy expended or speed. When choosing a cost function to use within our modelling, it’s common for defaults to prevail: for example with the use of the Time-based Tobler’s Hiking function.

When we plot these cost functions, however, we can see that they’re visually quite similar. For example, time-based cost functions on the bottom row mostly have this peak around zero with speed decreasing as slope increases.

The Campbell strava cost function on the bottom right, based a large sample size and a range of people of different sexes, ages, and fitness levels, however shows how the relationship between slope and speed can vary. And if we see this level of variability in the relationship, should we even be thinking about which cost function to choose? All cost functions approximate the relationship between slope and cost, and each contain their own biases.

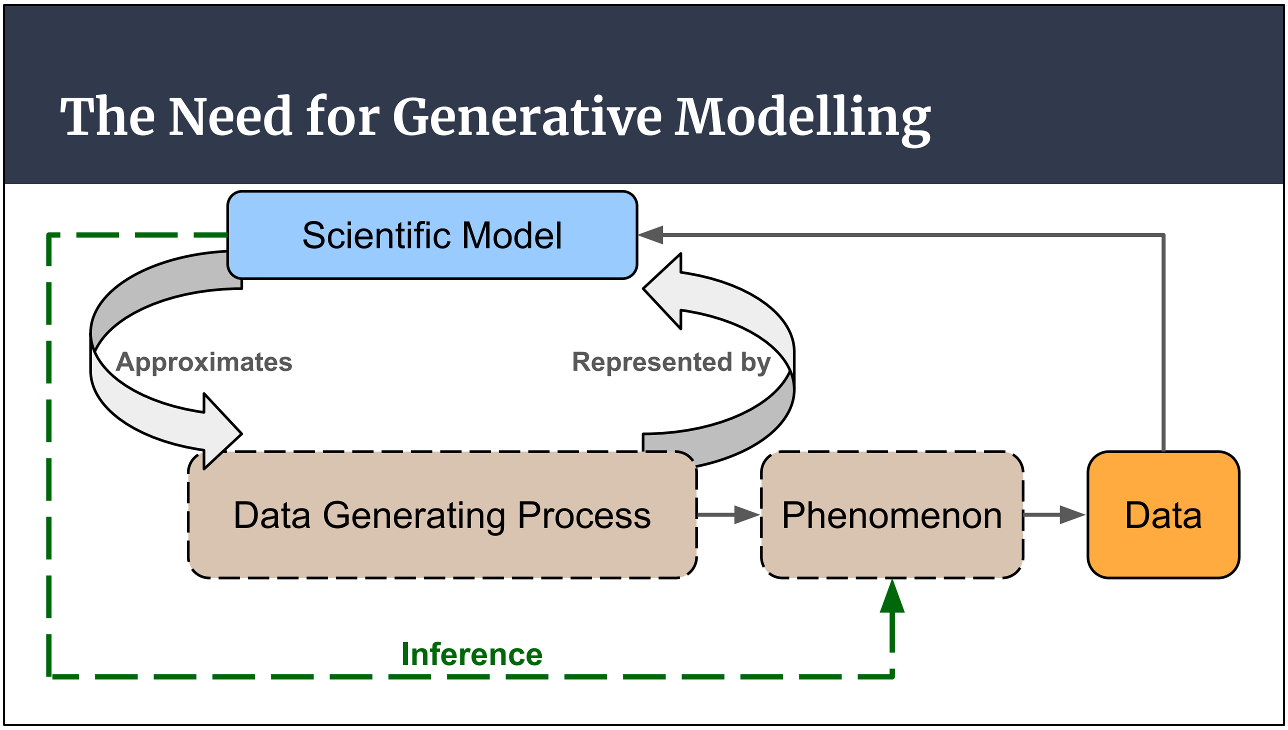
So, I argue that when aiming to explain known past routes, rather than having to choose a specific cost function, we should learn about the cost function from known past route that we’re modelling

Incorporating Additional Factors

And lastly, how do we incorporate factors other than slope into our models when aiming to explain known routes?

We know that factors other than slope influences movement, but we currently lack of a systematic approach for incorporating these within the models that we create.

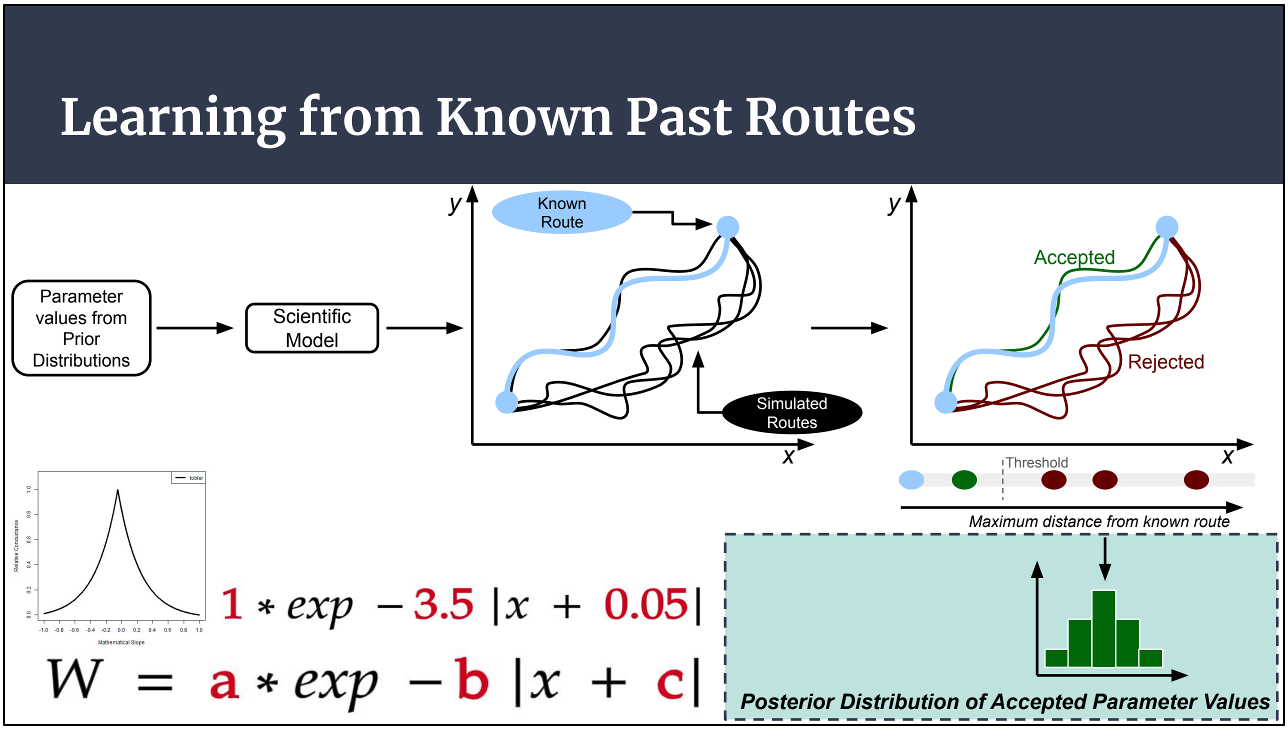
For this, I propose multi-criteria decision analysis: in short, this approach weighs multiple factors against one another when making decisions, with their resultant weights used as a proxy for preference and possible trade-offs



The Need for Generative Modelling

Before addressing how we operationalise this new approach, I also think we need to reconceptualise how we understand the relationship between modelling and explanation. When aiming to explain phenomena, we need to think about the data generating process - that is, the process that resulted in the phenomenon that we see in the world, or in short the story behind the data.

We can never access this data generating process nor the phenomenon directly but what we can do is collect data on this phenomenon. Using a scientific model that aims to approximate the data generating process we can work from the data that we collect back to the phenomenon. If the scientific model is able to sufficiently explain the data at hand, we can infer that the scientific model approximates the data generating process that created the phenomenon.



Learning from Known Past Routes

Parameter values from Prior Distributions

Scientific Model

Known Route

Simulated Routes

Accepted

Rejected

Threshold

Maximum distance from known route

Posterior Distribution of Accepted Parameter Values

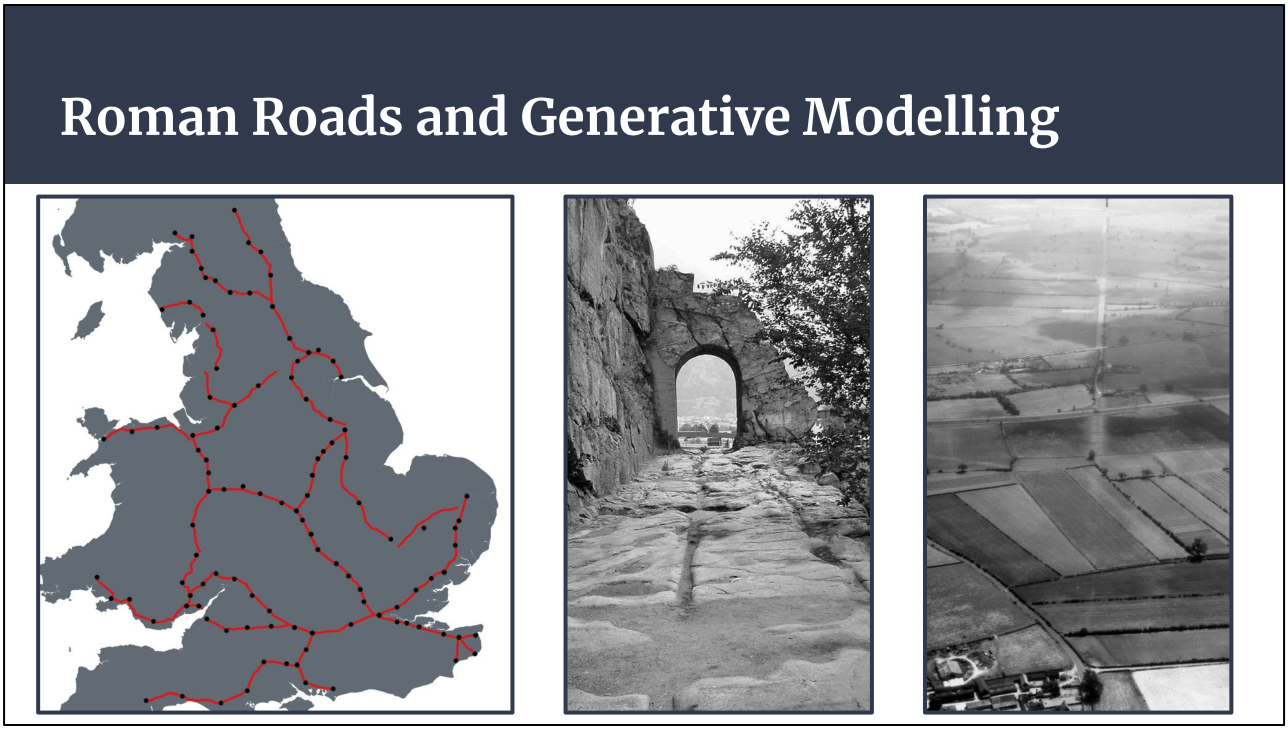
$$1 * exp\ -3.5\ |x\ +\ 0.05|$$

$$W\ =\ a * exp\ -b\ |x\ +\ c|$$

In light of generative modelling, how do we operationalise the new approach for explaining known past routes?

For this we can use Approximate Bayesian Computation, or ABC for short. Within the ABC framework, we have a scientific model with some parameters. For example, the double exponential function used in Tobler's Hiking function has three parameters, a,b,c. But unlike Tobler's Hiking function we're unsure what these values are. So we set what is called a prior distribution on each of these parameters. These are probability distributions reflecting our prior knowledge of what the parameter value could be before seeing the data

By inputting these parameter values into the scientific model we can produce simulated data. We then repeat this process with each simulation based on a different parameter value combination. These simulations are then compared against the known route that we're trying to model. If the distance between the simulated and known route is below a threshold value, we say that the simulated data is close enough to be deemed equal to the known route. We then take those accepted parameter values and create the posterior distribution: what the parameter values are more likely to be given the data that we're modelling.

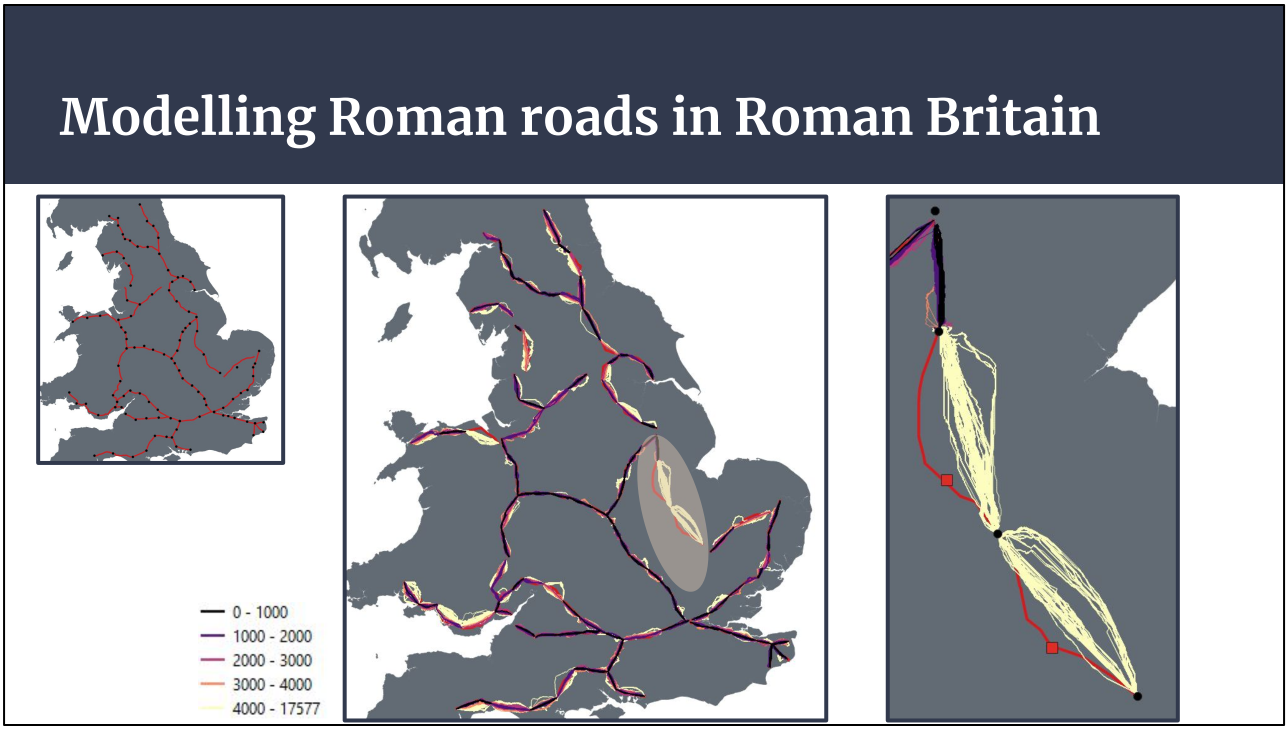


Roman Roads and Generative Modelling

Now for the case study. For this, I'll be aiming to explain all 114 Roman route sections in Britain that are recorded within the Antonine Itinerary (a late third century document of routes across the Roman Empire)

When thinking about the data generating process of Roman roads I've identified two factors that are deemed relevant for an initial model: slope and straightness. The factor slope is included to reflect that the course of Roman roads was influenced by slope. For example, based on a systematic analysis of Roman roads in Roman Britain, it is suggested that the target for the maximum slope was around 12.5 percent. Additionally, the factor straightness is included to reflect that where possible Roman roads were constructed in straight segments.

Whilst there are likely to be additional factors that are not included within this current model, such as the need to integrate pre-existing Iron Age sites, this conceptualisation provides an initial model that can be improved after deficiencies are identified



Modelling Roman roads in Roman Britain

Using the model just introduced and the ABC framework, the 114 Roman route sections in Britain were modelled. All the simulations are shown in the middle image and coloured by their distance from the corresponding route section. Each Roman road section is modelled independently, but hierarchical modelling is also possible

Looking at the simulations, we can identify that some of the routes were simulated within 1,000 metres of the known route section. This suggests that the current model of slope and straightness is sufficiently able to explain the data generating process of these sections

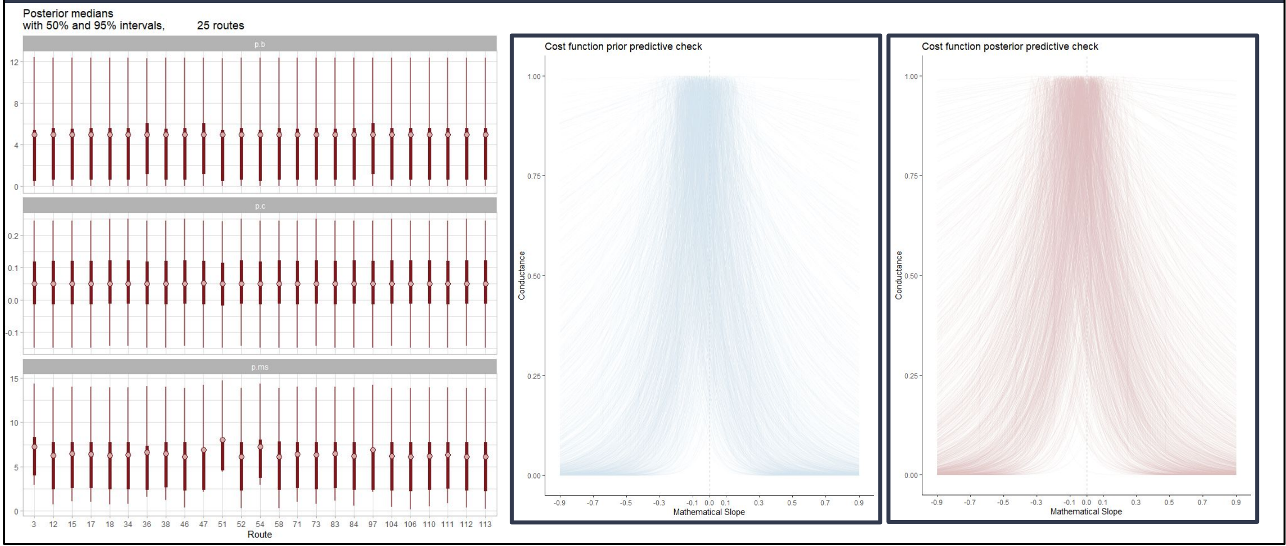
There are however some routes that we cannot adequately model. For example, the two road sections from Cambridge up to Lincoln, highlighted by the white oval. Looking at this particular case in more detail, the inability to adequately model the two routes is likely because of the Antonine Itinerary, on which the origin and destination of the routes is defined. More specifically, the Antonine Itinerary does not include two Roman forts constructed during the conquest of Roman Britain. These are shown by the the red squares. Given the dates of these forts, attributed to circa. AD 43 - 54, the forts would have influenced where the road sections were constructed. This suggests that current model is missing an important factor, namely forts

As a result, we can then expand our model incorporating our knowledge on potential factors that might have influenced where roads were constructed

**** ROMAN FORTS ARE GREAT CASTERON IN THE NORTH AND

GODMANCHESTER IN THE SOUTH

Posterior Distributions



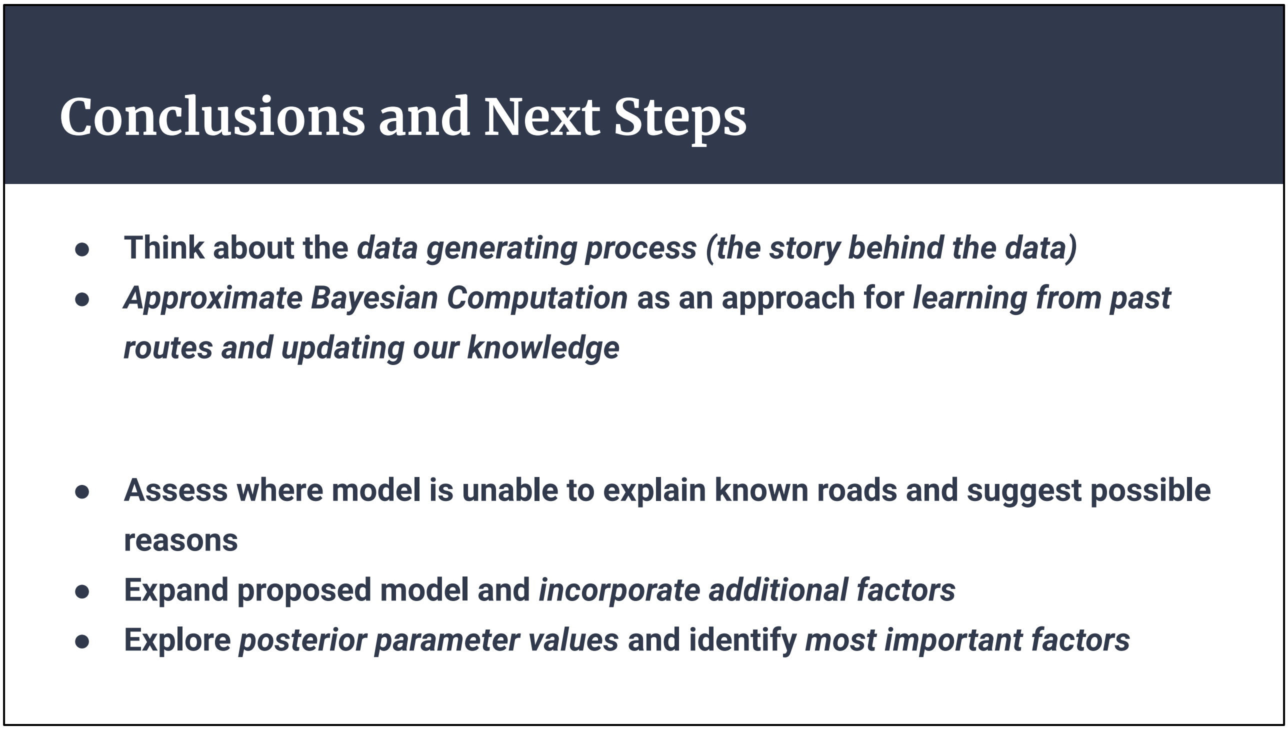
Despite the current model being unable to explain all of the Roman road sections sufficiently well, we can still inspect our posterior distributions, the parameter values given the data that we're modelling

Here we take simulations within 1,000m as sufficiently explaining the Roman road sections.

For example, looking at the b and c parameter values that are used to construct the cost function, we can see that there is little variability between each of the road sections. More variability is however present when looking at the maximum possible slope gradient which results in simulations that are able to explain each Roman road section.

Plotting the cost functions using the b and c prior and posterior distribution values for a single Roman road section, we can see that they're quite similar. This shows that the values that construct the cost function and represent the influence of topography on movement are not that important when aiming to explain the Roman road sections.

This suggests that a second model could remove the influence of the topography as measured by the cost function and instead include maximum slope gradient only



Conclusions and Next Steps

- Think about the *data generating process (the story behind the data)*
- *Approximate Bayesian Computation* as an approach for *learning from past routes and updating our knowledge*

- Assess where model is unable to explain known roads and suggest possible reasons
- Expand proposed model and *incorporate additional factors*
- Explore *posterior parameter values* and identify *most important factors*

Bringing it all together, two main conclusions can be drawn from this presentation:

(1) That to explain known routes, we need to think about the data generating process that resulted in the route that we're trying to explain. The models we use should reflect this data generating process, be descriptive in function, and grounded within the historical, political, and social dynamics in which the routes were created in

(2) Approximate Bayesian Computation is a viable method for incorporating prior knowledge on the factors influencing known past routes, and updating this knowledge given the known past routes that we're modelling

As my PhD is still ongoing, the next steps for my research will involve:

(1) First is to assess where the current model is unable to explain the Roman road sections and to suggest possible reasons for this

(2) Second is to iteratively modify the model and assess outputs against the Roman road sections. This will aim to create a better generative model that can explain more of the Roman road sections

(3) And lastly, to explore the posterior parameter values and identify which factors had most effect on where Roman roads were constructed. Spatial and temporal analyses of these posterior parameter values will also be undertaken

Thank you
Any questions?

Joseph Lewis
@josephlewis1992

CAMBRIDGE TRUST

Open-Oxford-Cambridge
Doctoral Training Partnership

UNIVERSITY OF
CAMBRIDGE

CAA2022 OXFORD
iN SIDE
FORMATION