ANNE Challenge - Written Summary of Challenge Process

This document (based on the ANNE template) is part of my ANNE challenge submission alongside the shared GitHub repository and final report “**Environmental-impact-of-housing-in-england.pdf**”.

Challenge Name: ANNE Analysis Challenge

GitHub Repository: https://github.com/nclJoshCowley/ANNE-challenge

Candidate Name: Josh Cowley (j.cowley1@ncl.ac.uk and josh.cowley@hotmail.com).

> **Do you give consent for NEECCo to contact you in future about your challenge submission?**

Yes, feel free to contact me at the emails above

> **How and why did you decide to pursue your chosen idea? What other ideas did you consider? (Details on the brainstorming process)**

More information is available at the attached GitHub repository (see IDEAS.md).

My initial approach was concept-first, so I came up with two ideas described below, before looking at data availability.

Idea 1: spatiotemporal modelling on phosphorous (by product of agriculture runoff and sewage) in water catchments.

Idea 2: modelling EPC with the associated *improvement* shown on EPC certificates.

Decided against water idea as I could not find high quality data on the topic, housing idea changed throughout the process to fit a more data-first approach. That is, the concept-first approach did not work as hoped.

> **Can you briefly describe the data and analytical methods that you considered and used in your work, including any limitations you came across and how you overcame these?**

Data. I came across the Local Authority Housing Statistics (LAHS) in my initial research, upon emailing the relevant persons about a more granular version of these data the English Housing Survey was highlighted. My project uses only one data source (EHS) but more would be better.

Methods. Linear models are abundant and the multilevel modelling has surfaced in my PhD research where group levels are known (for my research application, group levels are inferred).

Limitations. The standout limitation is the single data source and using the EPC/SAP ratings, the analysis will only be as good as these ratings, more causal data would improve things.

> **Can you summarise the main results from your analysis and the impact that these findings have in the wider context (with regards to the work of NEECCo)?**

Main result (Figure 3.1) is average house efficiency is on the increase, as with many small projects this asks more questions than it answers. So, we (and NEECo) should ask is this increase substantial? And what is the impact of this increase?

Figure 3.2 claims that the North East has marginally more efficient housing than other regions which could be a key talking point for NEECo.

During research phase, we found a call for analysis of EPC, this model could contribute to that given more detailed dwelling data.

> **If you were able to continue working in this area, what would be your next steps? Is there any future development you would like to work on? Can you see any challenges with this?**

Again, using these models with more detailed dwelling data would be useful as current predictors are known influences of EPC. Wider variable selection could highlight dwelling aspects that aren’t correlated with EPC that perhaps should be.

Analysis could be done at population level (model A, linear model) or at a group level (model B, multilevel model) depending on the questions one hopes to answer.

With the EHS data, we could use the interview part of the dataset to further understand fuel poverty, which would have a greater impact.

> **Can you give an example of analysis you have performed to investigate a problem, prior to this challenge?**

I am currently researching groundwater monitoring of hydrocarbons, say the carcinogen benzene, using telemetry such as pH, temperature, dissolved oxygen etc.

We have tried various exploratory and modelling techniques such as matrix normal distribution (multiple correlated response variables) and currently looking into mixture of experts.

Outside of my PhD research I have aided the analysis of patient data for advanced laryngeal cancer using Cox regression.

> **Was there a part of this challenge that you found particularly difficult? How did you overcome this?**

A lot of my time was spent in the problem understanding phase (with data understanding for preliminary datasets). So time management became a real issue as I had underestimated the time required for the challenge overall.

Thankfully, I found a high quality and interesting dataset and focused on that decreasing the scope of the project to ensure a submission could be done in a timely manner.

> **What is one skill you feel you have developed through this challenge and how do you think this will benefit you in future work/study?**

This is the first project where I have been actively thinking about design and aesthetic of the overall project as a mean to communicate more effectively.

However, in light of my answer to the previous question, the one skill developed is likely project management (specifically data science projects). This is in no small part thanks to the workshop on workflows highlighting methodologies such as CRISP-DM.