Smith institute

Machine Learning Discovery of Vulnerability Signatures

Document SI22-P22-001

Version 1.1

8 November 2022

Contract 1222037571

Authors	**Kieran Kalair**
	Alex Bowring
	Cameron Booker
Internal reviewers	**Rachael Warrington**
	Francis Woodhouse

Executive Summary

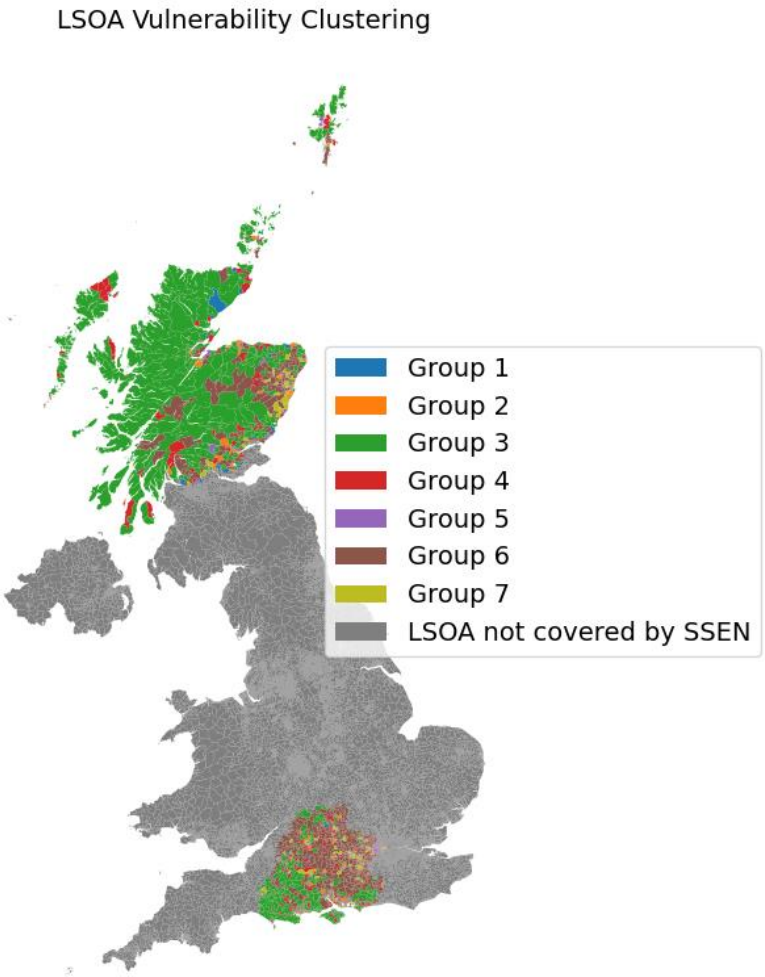
No vulnerable customer should be left behind during the UK's transition to net-zero. To ensure that future network investment decisions support this aim, Scottish & Southern Electricity Networks (SSEN) need insight into the underlying drivers of vulnerability in the different locations they serve, and into the commonalities and differences in these drivers across locations. That insight is exactly what this project provides, through mathematics and machine learning.

We identify seven distinct groups of locations that SSEN service, with each group having similar demographic features driving vulnerability, and each requiring distinct network investment strategies to ensure that those who would be particularly adversely affected by the risk of frequent power cuts will be served by secure, stable, and resilient future networks. The methodology and results of this study can also inform wider action and investment by government and other organisations to address the underlying drivers of vulnerability. These seven groups cover different locations across Southern England and Scotland: a single strategy for investing in network infrastructure in Scotland and another for the south of England is therefore insufficient to address the varying causes of vulnerability that customers experience. SSEN will need network investment strategies that are geographically tailored to provide resilient networks in a context of different locational drivers of vulnerability.

To identify groups with common drivers of vulnerability, we perform a mathematical analysis of demographic data and of SSEN's priority service register, using machine learning to model the complex relationships that exist between the two. We then use these complex relationships to understand what drives vulnerability in each location that SSEN service. Finally, we find natural groupings of vulnerability drivers using these results, attaining a mathematically sound and data-driven understanding of the vulnerability landscape across SSEN's areas of operation. Note that we focus our analysis on areas that SSEN service, and the findings apply only to those areas, but the methodology is transferable to other geographical areas, other Distribution Network Operators, and other utility providers. For locations that SSEN service, the seven distinct groups that share common vulnerability drivers are detailed below, followed by a depiction of their geographic locations.

Group Number & Level of Vulnerability	Description of Group
1 – Very high	Driven up by higher levels of poor health and disability/mental health benefit claimants, reduced by smaller household sizes.
2 – High	Driven up by larger household sizes, reduced by lower elderly population levels.
3 – High	Driven up by larger elderly population levels, reduced by lower levels of disability and mental health benefit claimants.
4 – Slightly higher than average	Driven up by larger elder population levels and moderately higher provision of care, reduced by smaller household sizes.
5 – Slightly lower than average	Driven down by lower elderly population levels and larger levels of ethnic diversity, increased by higher household sizes and greater provision of care.

6 – Low	Driven down by lower level of bad health and disability/mental health benefit claimants, increased by moderate elderly population levels and household sizes.
7 – Very low	Driven down by substantially lower elderly population levels, less provision of care and a higher level of households in private rented dwellings.



Using the identified groupings, SSEN can ensure future investments in developing the network truly account for the levels of vulnerability their customers in different locations face. Targeted interventions can be made to ensure that the network is secure, stable, and resilient in the areas where vulnerable customers would be most disadvantaged by more frequent power cuts, providing tailored investment that supports the communities SSEN serve.

The analysis contained within this report lays the foundation for a new type of energy scenario: vulnerability future energy scenarios, which can be used by Distribution Network Operators to ensure their future network investments aid the communities they serve and leave no vulnerable customer behind. Future work could use mathematical and machine learning techniques to generate robust forecasts at the geographic level required for these vulnerability future energy scenarios.

Contents

Document History

Version 1.0.1	27 October 2022:	Initial issue
Version 1.1	8 November 2022	Minor revisions responding to SSEN feedback

1 Introduction

The energy sector is undergoing a fundamental shift in its operation and capabilities, with the introduction of renewable energy, a drive towards net-zero and a cost-of-living crisis catalysed by rising wholesale energy prices. Confronted by a rapidly changing energy landscape, Scottish & Southern Electricity Networks (SSEN) must make infrastructural investments to ensure the distribution network remains fit for purpose over the upcoming decades.

In the context of network management, it is especially important that SSEN's most vulnerable customers continue to be supported. However, to act effectively it is vital that SSEN first understand what vulnerability looks like from a demographic perspective. While a human may be able to identify individual characteristics that drive vulnerability through manual analysis, the circumstances that lead to a customer's inclusion on the Priority Services Register (PSR) can vary widely and are likely to involve many interacting factors. At the level of neighbourhoods – the communities SSEN service – the nature of vulnerability becomes even more complex: what is the relationship between a community's demographic make-up and its associated level of vulnerability? Answering this question will allow SSEN to ensure that their network investment decisions support vulnerable customers in the future, while also providing insight that can help SSEN make enhanced decisions to help their most vulnerable customers in the short-term.

In this report, mathematics and machine learning are applied in conjunction with data and expert advice provided by SSEN to understand the common demographic characteristics that drive consumer vulnerability. First, we develop a model to predict vulnerability in each of the Lower Layer Super Output Areas (LSOAs) that SSEN serve, based on each LSOA's demographic make-up. Following this, we apply cutting-edge explanatory methods to obtain 'vulnerability signatures': groupings of LSOAs whose vulnerability levels are driven by a set of common factors. The benefits of this are two-fold. First, the vulnerability signatures will not only identify the most vulnerable communities, but also uncover the underlying common factors leading to higher levels of vulnerability. This will support SSEN in determining the societal areas where investment is best placed now. Second, these groupings will also highlight both geographic and demographic areas that may need more attention from SSEN in the future, facilitating further forecasting of vulnerability scenarios.

This report complements parallel work being conducted by Imperial Consultants (ICON) and National Energy Action (NEA) to assess vulnerability. While we use a data-driven mathematical analysis to identify the specific demographic factors causing vulnerability in the areas SSEN service now, ICON are applying foresighting approaches to provide a broader view of future changes in vulnerability.

This work to assess vulnerability to support investment decisions in Future Energy Scenarios supports SSEN's strategies contributing to the UN Sustainable Development Goals:

- Goal 7: Affordable and Clean Energy to All - this work aids SSEN's investments to ensure affordable and clean energy to all, as the energy landscape is changing across Scotland the England.

- Goal 9: Build resilient infrastructure – The objectives in this report will aid investment decisions increasing the resilience and availability of energy in LSOAs.
- Goal 10: Reduced Inequalities - the work helps reduce inequality through ensuring investments account for the most vulnerable consumers.
- Goal 11: Sustainable Cities and Communities – the insights provided by this work support investments towards sustainable cities and communities.
- Goal 17: Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development – This project has been realised by collaboration between parties to develop a technology solution with wider and global application.

The remainder of this report is structured as follows. We first detail the dataset provided, and then discuss the technical steps required for this project. The technical steps constitute:

- Building a model of vulnerability that takes demographic data for locations SSEN service as inputs, and outputs a prediction of each location's vulnerability level
- Generating explanations from this model, to understand why it is predicting the vulnerability values it provides
- Identifying locations with drivers of vulnerability that are similar to one another and interpreting different groups that arise in the data

We then discuss the results obtained from following these technical steps, and end by summarising how SSEN can use these conclusions, including detailing recommendations for future studies.

1.1 Data Summary

SSEN provided Smith Institute with demographic data for all 6009 LSOAs that the company serve. A single LSOA represents a specific spatial region of Great Britain, and SSEN's distribution network contains regions in Scotland and the south of England. We are aware that the term 'Datazones' may be used to refer to locations in Scotland, but for consistency throughout the report we will use the term LSOA to apply both to locations in Scotland and the South of England. The data covers 43 factors that may impact the total vulnerability within each LSOA, including metrics related to population age, health benefits, income, social isolation, internet usage, housing, and qualification details. These metrics represent the ***features*** or ***inputs*** that our model building process considers, as described in section 0.

1.1.1 Computing vulnerability

Vulnerability in the Great Britain is measured at the household level and recorded on the Priority Services Register (PSR). Members of the public may join the PSR under a range of conditions that include having a disability, relying on electrically powered medical equipment and being of pensionable age. For this project, discussions with SSEN led to a definition of vulnerability within each LSOA, computed as ***the fraction of households in the LSOA that have at least one resident on the PSR.*** Therefore, vulnerability is measured between 0 and 1 throughout this report, with higher values indicating LSOAs that have a large fraction of households with residents on the PSR.

It is important to note that it is up to individual consumers to apply to be added to the PSR through appropriate means. It is therefore widely believed that the actual number of households on the PSR is below the total number of households that would be eligible for inclusion. Because of this, SSEN have developed a methodology to estimate the total level of vulnerability for each LSOA, including *all* households that would be eligible to be added to the PSR. However, through discussions with SSEN, Smith Institute learned that this estimate incorporates the same demographic data provided for this project. Therefore, if this estimate was used to represent vulnerability in this work, our machine learning model may simply re-learn the model that was used by SSEN to generate their vulnerability estimate. To avoid this, we instead use ***the actual number of households on the PSR*** when computing LSOA vulnerability. Our model therefore captures the inherent drivers of vulnerability for all households currently registered on the PSR. We show the distribution of this vulnerability metric in Figure 1.

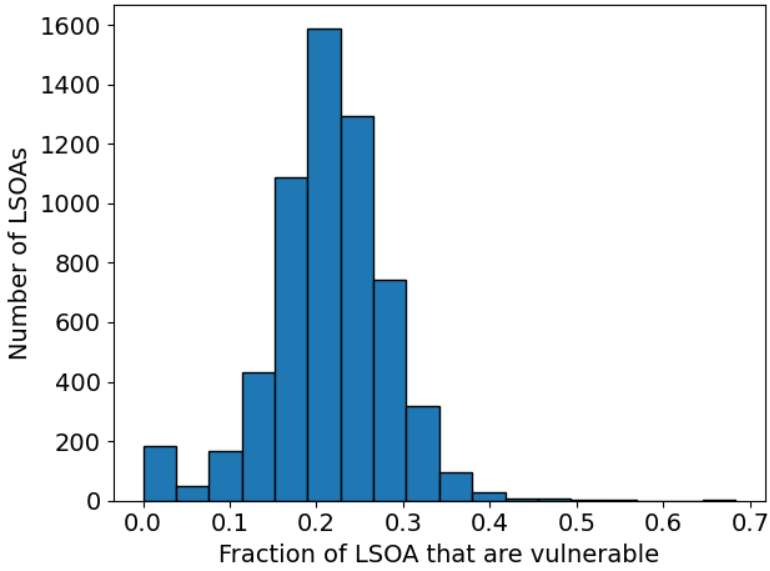


Figure 1: Distribution of vulnerability in LSOAs that SSEN service. Typical vulnerability lies just above 0.2, meaning that it is most common to observe around 20% of households in an LSOA having an occupant on the PSR. We see the highest levels of vulnerability reach just below 0.7, whereas the lowest are close to 0.0. These LSOAs with very low vulnerability are in fact an artefact of the data collection process, discussed at length in section 2.3.4.

1.1.2 Additional data used for visualisation of outputs

Since LSOAs close to each other are more likely to have a similar demographical make-up, a further question addressed by our analysis is *where* in Great Britain is investment most needed? To help answer this question, we use additional open-source geographic data to visualise the outputs of our analysis spatially across Great Britain. The relevant data is collected from the 2011 census[1], and using this, outputs from this project are overlaid on a map to highlight the physical locations of LSOAs with similar drivers of vulnerability. To provide an example of this data, the vulnerability of each LSOA in SSEN's network is visualised in Figure 2. Our analyses will provide

[1] Shape file data can be downloaded from 2011 Census Geography boundaries (Lower Layer Super Output Areas and Data Zones) - Dataset - UK Data Service CKAN

further inference on these raw vulnerability figures, shedding light on the factors that drive vulnerability in each of these LSOAs, and which LSOAs share common drivers of vulnerability.

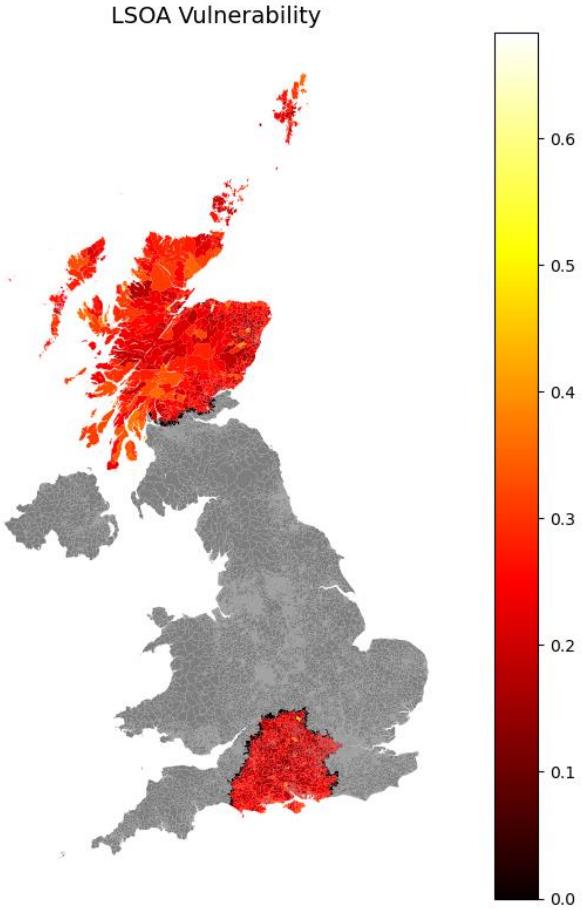


Figure 2: Visualisation of vulnerability across the UK. LSOAs that SSEN service are coloured corresponding to their vulnerability, with vulnerability increasing as we move from black to red to yellow to white colours. LSOAs that SSEN do not service are marked in grey, and we have no associated data on these. Note that the LSOAs with very low vulnerability (around 0.0) are in fact an artefact of the data collection process, discussed at length in section 2.3.4.

2 Methodology

2.1 Technical workflow

The technical steps required to perform the analysis for this project are detailed in Figure 3, showing a depiction of the analysis pipeline that takes data provided by SSEN and generates insights. In the subsequent text, we detail why each of these steps is required.

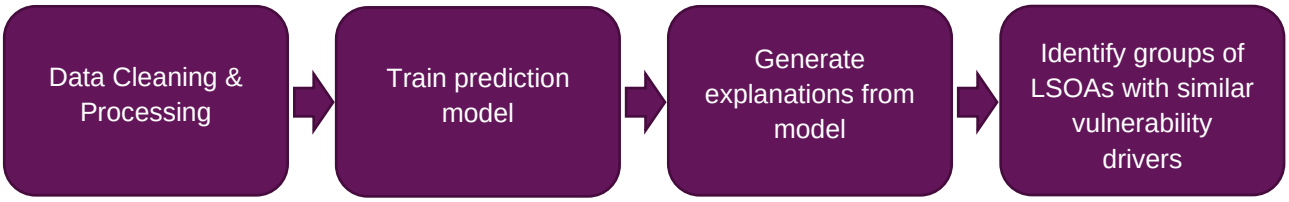


Figure 3: Analysis pipeline required to take demographic data for each LSOA and discover groups of LSOAs that have similar vulnerability drivers.

First, pre-processing is applied to the data to ensure it is in a format appropriate for modelling and analysis. We then train a prediction model that takes as inputs demographic data for a given LSOA, and outputs a prediction of that LSOA's vulnerability. It is important to note that we are not building this prediction model as a method for determining future vulnerability: we are instead using it to identify patterns and relationships in the provided data that relate LSOA demographic values and vulnerability. After training the prediction model, we then generate explanations from it, telling us the specific demographic characters most influential in determining the model's vulnerability estimate for each LSOA. The explanations methodology is detailed in section 2.4. After obtaining these explanations, we finally identify groups of LSOAs that have similar drivers of vulnerability, yielding a data driven assessment of what drives vulnerability in the LSOAs SSEN operate in, and what groups of LSOAs exist with similar drivers of vulnerability.

2.2 Pre-processing of data

The data SSEN provide for this project offers a wide range of demographic features for LSOAs across Scotland and the south of England. We apply pre-processing steps to the data to prepare it for modelling and analysis. We:

- Generate an age field 'between 65 and 85'. In the original data, old-age fields were split into three categories: 'over 65 years old', 'over 75 years old' and 'over 85 years old'. These fields are highly dependent on each other: as the number of over 75's increases, so must the number of over 65's. To simplify the combination of age features, we generated a feature measuring the proportion of people within an LSOA's population aged between 65 and 85. We did the same for the population between ages 5 and 16.
- Encode the field 'Internet user engagement type' from strings (for example 'passive and uncommitted users' and 'youthful urban fringe') into integers (1, 2, …, 10), as this allows them to be used as features by the prediction models.
- Combine the two separate columns are given detailing fuel poor households – one with entries for LSOAs in Scotland and one with entries for LSOAs in the south – into a single column.
- Generate a feature 'average number of persons per household' as it may be insightful as a measure of population density and its impact on vulnerability.
- Remove fields not appropriate for modelling:
    - The estimate of PSR eligibility (see discussion in section 1.1.1)
    - Age fields made redundant from the additional fields discussed above
    - Deciles of distances to services
    - The fields 'network investment priority score', 'community resilience score' and 'social isolation score'. These three scores are derived using demographic data available for each LSOA; for example, the social isolation score used information about the number of lone parents and single pensioners. Since it is not the scores that will drive vulnerability in each region per se, but the underlying demographical

features from which the scores were derived, the three scores are removed from the data.

Finally, a process called ‘one-hot encoding’ is applied to the features ‘fuel poverty levels’ and ‘internet user engagement type’. This ensures that the model can effectively utilise both features that have distinct categories rather than continuous values. In the example of internet user engagement type, it does so by replacing the original feature and its 10 possible categories with 9 one-hot encoded features: ‘internet user engagement type 2’, ‘internet user engagement type 3’, …, ‘internet user engagement type 10’ where each of these features takes the value 0 or 1, depending on which internet user engagement type an LSOA has. Only 9 one-hot encoded features are needed to replace the original 10 categories as, when all 9 of the one-hot encoded features take the value 0, the model can deduce that the internet user engagement type for an LSOA must be 1.

2.3 Building a model of vulnerability

Having processed and cleaned the data, we turn to building a model that predicts vulnerability in each LSOA. We wish to identify the complex relationships that exist between an LSOA’s demographic features and its vulnerability. In this section we detail what form this model takes, what features it uses and its performance on the provided data. We then describe the explanation methodology applied to the fitted model, from which we identify the main demographic factors that influence the model’s vulnerability predictions.

2.3.1 Model selection

To build a model that predicts vulnerability, we consider two candidate approaches. The first is a simple linear model, known as an ‘elastic net’ in the wider literature. The elastic net predicts vulnerability as a linear combination of demographic predictor features, for example:

$$vulnerability = \beta_0 + \beta_1 \times Disability\ benefits + \beta_2 \times Average\ household\ size$$

Here, two features – disability benefits and average household size – have been used for demonstrative purposes, while in practice many more features would be considered. In this example, the model fitting procedure identifies the optimal values of β_0, β_1 and β_2 such that the model most accurately matches the data.

One advantage of the elastic net model is that it can filter out redundant features and focus only on those that provide the most useful predictive performance. This is beneficial for our application, where SSEN have provided data on tens of features. Considering more than one model also allows us to assess what we gain in terms of model performance as the model complexity is increased. However, a drawback of the elastic net is that, by virtue of being linear, it is unable to capture all the complex (and likely nonlinear) relationships that exist in the data.

An alternative modelling approach that has shown state-of-the-art performance on tabular data is known as a ‘gradient boosted tree’. This model uses a collection of decision trees to predict vulnerability. Two example decision trees are pictured in Figure 4. Such a model takes a given LSOA and its associated demographic features as an input and evaluates a set of criteria to

determine which ‘leaf’ of the decision trees the LSOA falls into, and hence what prediction of vulnerability is returned.

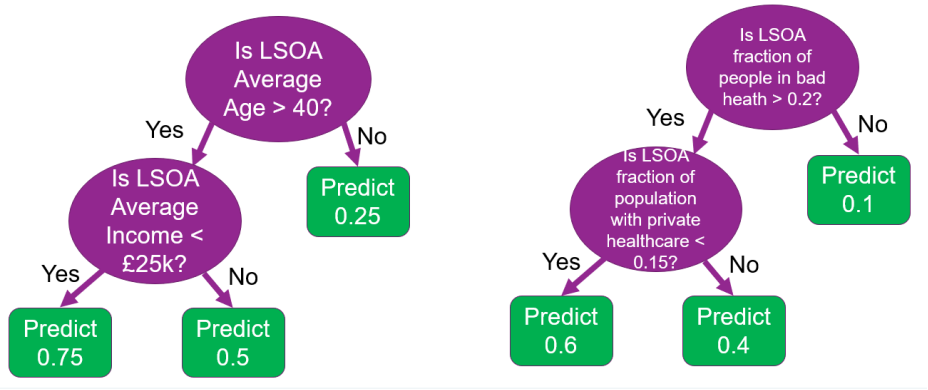


Figure 4: An example of two distinct decision trees one could build using the LSOA demographic data. The first (left) considers the average age and income and uses these to generate a prediction of vulnerability. The second (right) considers the population’s health and access to private healthcare. There are illustrative trees to demonstrate the concept, not ones used by the final model.

A gradient boosted tree builds upon a decision tree in several ways. First, rather than using a single decision tree to generate predictions, it combines the results from many different trees. Second, trees are constructed such that each tree successively improves predictions made from the previous set. To do this, an initial tree is trained to predict vulnerability given certain demographic features of an LSOA. This prediction will not be perfect, and there will be an associated error (the difference between the true vulnerability and the vulnerability predicted by the model). To reduce this error another tree is added, using a potentially different set of demographic information to the first tree to gain a better prediction. This process is repeated successively, generating a set of trees that incrementally improve the prediction accuracy.

We apply both an elastic net and gradient boosted tree to the LSOA demographic data and show comparisons of their performance in section 2.3.5. Due to the superior performance of the gradient boosted tree model, capturing more complex relationships present in the data, we use this as our model to generate explanations from in the future sections. We did not consider alternative architectures due to the tabular nature of the data, which is often well modelled by tree-based methods, and also for the reduced computation time when generating explanations (see section 2.4).

2.3.2 Hyperparameter Tuning

Most machine learning models, including those used in this project, require the modeller to select two things before they can be used. The first is a set of ***features*** that the model can use. These are the demographic details for a given LSOA, for example the age of the population and levels of income. The second is a set of ***hyperparameters*** that inform how the model can learn relationships in the data. In the context of a gradient boosted tree model, these hyperparameters include the number of layers the tree can have, and the total number of trees that are combined to reach a vulnerability prediction. To determine these hyperparameters, we use a method called cross validation, in which the data is split into several partitions. Suppose we choose 5 partitions: we then train the model on 4 of these and evaluate its performance on the remaining partition. This

is repeated but changing which partition of data the model performance is evaluated on, until each partition has been used as the evaluation partition. This is repeated many times for different hyperparameter choices, and the final set of hyperparameters is chosen as the set that minimise the average evaluation error. A visual depiction of cross-validation is given in Figure 5.

	Data Record 1	Data Record 2	Data Record 3	Data Record 4	Data Record 5	Data Record 6	Data Record 7	Data Record 8
K=1								
K=2								
K=3								
K=4								

Figure 5: Visual example of cross validation, with 8 data records and 4 folds used. Data that is used to train the model is shown in blue, whereas data that is used to evaluate performance on each fold is shown in green. As we move from folds 1, 2, 3 to 4, we train the model with a different subset of data and evaluate it on the remaining data. The final performance of a given set of hyper-parameters will then be computed as the average across the green data records when they are withheld from training.

It is important to note that selecting hyperparameters through cross validation avoids introducing bias into our evaluation. If we were to use the entire dataset when determining the optimal hyperparameters, we would have less reliable estimates of the model performance for a given hyperparameter set, as we would be fitting and evaluating the model on the same dataset, which would then hide any issues where the model is too specific to the given data and relying on noise rather than meaningful relationships.

2.3.3 Feature selection

2.3.3.1 Refining the set of features the model can use

Bespoke model building in machine learning requires several careful choices. The first is the choice of model (discussed in section 2.3.1) and the second is the choice of hyperparameters to use (discussed in section 2.3.2). A further key consideration is what demographic features should the model use to make vulnerability predictions. A naive approach might be to include every feature at our disposal; however, this can lead to difficulties in fitting the model. A specific problem, known as multicollinearity, can occur when multiple features the model uses are too similar to each other. When two demographic features are highly correlated, the model can find it difficult to distinguish the individual effects of each given feature on vulnerability. This could lead to imprecise and unrobust effect estimates, which may distort our conclusions about how each demographic factor influences vulnerability. To mitigate this problem, we evaluate the similarities between all feature data available, and with expert guidance from SSEN, only keeping the most appropriate feature when multiple features are found to be highly similar. To assess if the features available for

use in this project are highly correlated, we compute the correlation coefficient between all possible pairs of features and show results in Figure 6.

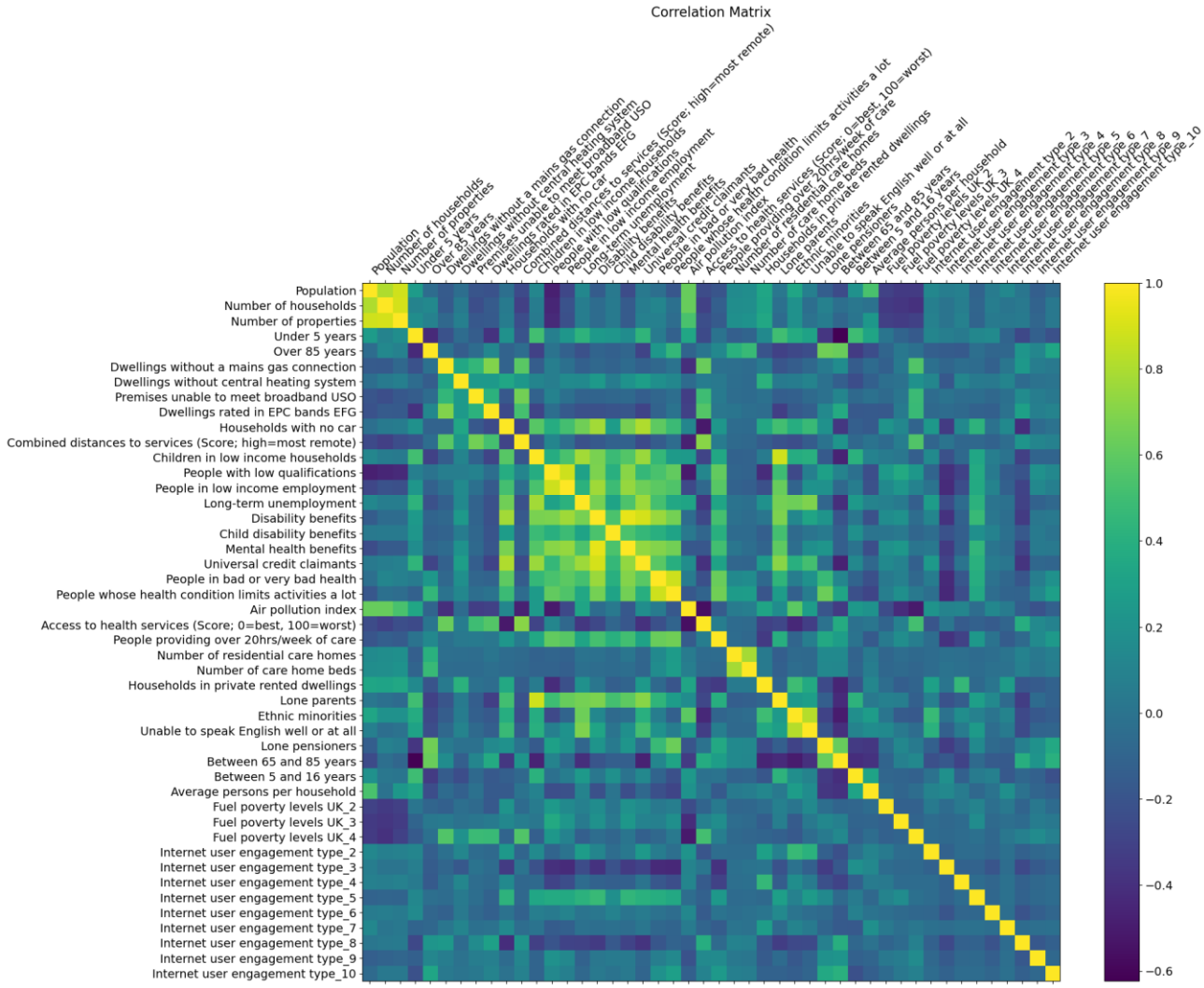


Figure 6: Correlations between all features that are available to use in this project for LSOAs. Yellow indicates a very high positive correlation, and hence strong relationship between the two features – as one feature increases, so does the other. Dark blue indicates a very high negative correlation – as one feature increases, the other decreases.

From Figure 6, we see the following highly correlated feature subsets:

- ‘Population’, ‘Number of households’ and ‘Number of properties’: we only include number of households for use with the model, omitting the other two.
- ‘People with low qualifications’ and ‘People in low-income employment’: we only retain people in low-income employment for use with the models.
- ‘People in bad or very bad health’ and ‘People whose health condition limits activity a lot’: we only consider people in bad or very bad health for use with the models.
- ‘Number of residential care homes’ and ‘Number of care home beds’: we only retain number of residential care homes for use with the models.
- ‘Ethnic minorities’ and ‘Unable to speak English well or at all’: we retain ethnic minorities for use with the models. This was discussed at length with SSEN and one could select

either of these two features, however to ensure a robust model is built, one must ultimately make a choice between these two.

- 'Lone pensioners' and 'Between 65 and 85 years old': we only consider between 65 and 85 years old for use with the models.

While the correlation between these features does indicate they are highly related in this dataset, we understand that they are not identical. In each of these cases, it is important to note that the retained feature is likely to represent a combination of itself and its corresponding correlated features in our analysis. For example, while the 'Ethnic minorities' predictor has been retained following our correlation evaluations and discussions with SSEN, the contribution of this feature to our modelling and subsequent explanations should be viewed as a combination of the individual effects attributable to both the 'Ethnic minorities' and 'Unable to speak English well or at all' features. To better isolate the individual effects for groups of correlated predictors, SSEN could consider collecting further data on these specific characteristics and utilising statistical techniques robust to correlations as part of follow-up analyses.

Having refined the set of features as described above, we then do a final sweep of the remaining candidate features to remove any that may not be informative to the model. To do this, we fit an initial gradient boosted tree model and evaluate how many times each variable is being used as an evaluation criterion in the associated trees. An unused or rarely used variable indicates that the information the variable provides to the model is already captured in the other available features, or that the variable simply doesn't have a causal effect on vulnerability. Doing this process of feature removal identifies 'number of residential care homes' as a feature with little importance, and so we remove this from the set of features the final models can use. It is important to note that this model is not the final gradient boosted tree used, rather it is just a preliminary model used to assess in a broad sense feature importance. The final model used undergoes further tuning to ensure optimal performance.

2.3.3.2 Observations on air pollution index

One candidate feature that is provided in the data is 'air pollution index'. However, preliminary analysis indicates that inclusion of this feature in the model could be problematic. To demonstrate why, we show the distribution of air pollution index in Figure 7. From the figure, it can be seen that the air pollution index broadly indicates whether an LSOA is in Scotland or England; in general, LSOAs in the south are more likely to have a higher air pollution index then those in the north. Due to this discrepancy, the model may use the air pollution index as a proxy to explain how more fundamental differences between England and Scotland lead to differences in vulnerability. For example, since our feature data does not include country-specific economic or governmental policy information, the model may wrongly assign the effect these differences have on vulnerability to air pollution. Ultimately, this could lead to a misleading and exaggerated interpretation concerning the effect of air pollution on vulnerability, and therefore we remove air pollution index as a feature. In addition, we also include a binary 'Scotland_yn' nuisance variable, which takes a value of 1 when an LSOA is in Scotland, and 0 otherwise, to mitigate the effect of country-specific differences not captured by the data in our findings.

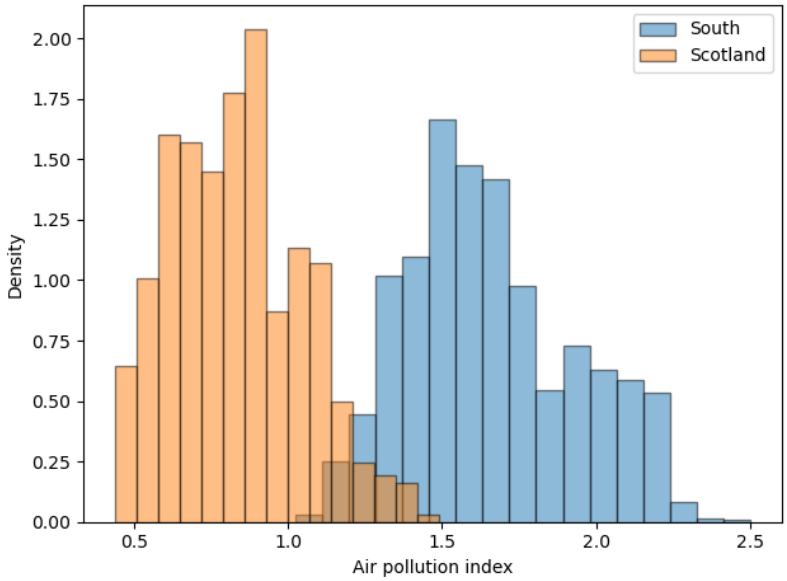


Figure 7: Histogram of air pollution index, separated into LSOAs in Scotland (orange) and the south (blue). The air pollution index in the two locations is vastly different. The LSOAs with highest air pollution index in Scotland are comparable to the LSOAs with lowest air pollution index in the south.

2.3.3.3 Features available to the model after selection

The full list of features the model can use is detailed in Table 1.

Feature	Additional Notes
Population	
Under 5 years	
Over 85 years	
Dwellings without a mains gas connection	
Dwellings without central heating system	
Premises unable to meet broadband USO	USO is 'universal service obligation', a UK wide measure.
Internet user engagement type	Engagement types given in the data as: Passive and Uncommitted Users Youthful Urban Fringe e-Veterans e-Professionals e-Withdrawn e-Mainstream e-Cultural Creators e-Rational Utilitarians Digital Seniors Settled Offline Communities
Dwellings rated in EPC bands EFG	
Households with no car	

Combined distances to services (Score; high=most remote)	
Children in low income households	
People in low income employment	
Long-term unemployment	
Disability benefits	
Child disability benefits	
Mental health benefits	
Universal credit claimants	
People in bad or very bad health	
Access to health services (Score; 0=best, 100=worst)	
People providing over 20hrs/week of care	
Households in private rented dwellings	
Lone parents	
Ethnic minorities	
Fuel poverty levels UK	
Between 65 and 85 years	
Between 5 and 16 years	
Average persons per household	
Scotland_yn	

Table 1: List of features that the models can use, after feature selection and refinement has been performed to isolate a subset of features appropriate for modelling.

2.3.4 Accounting for LSOAs that are not fully serviced by SSEN

Following discussions with SSEN, we found out that there are several LSOAs where energy network operation is shared between SSEN and other DNOs. Because of this, SSEN only service a subset of the total households in these LSOAs. However, the vulnerability data provided for this project is calculated as the ratio of households in SSEN's network included on the PSR to the *total* number of households in each given LSOA. Therefore, for the LSOAs shared between SSEN and other providers, these computations can vastly underestimate of the true levels of vulnerability. This can be seen in Figure 8, where the 'border LSOAs' surrounding the south and Scotland appear black due to having substantially lower vulnerability metrics, owing to this issue of only being part-covered by SSEN's network.

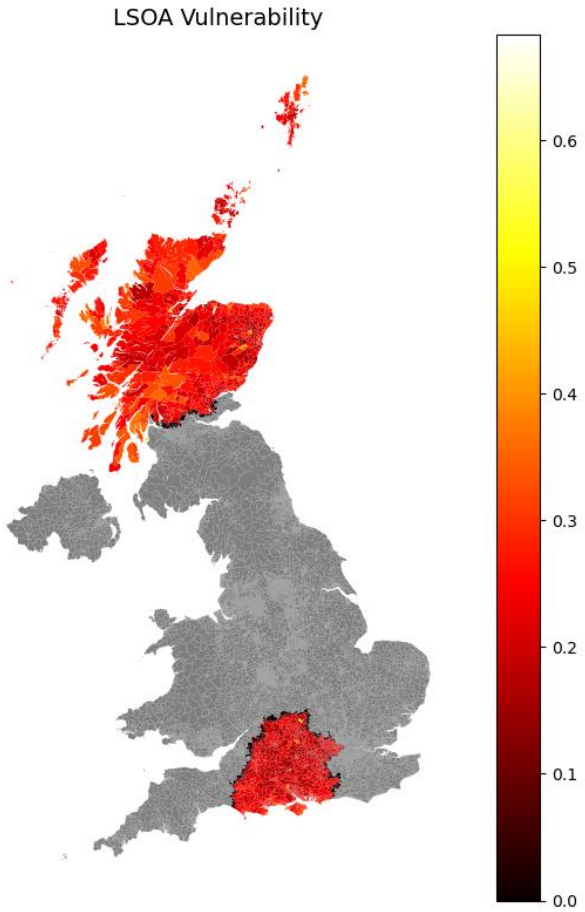


Figure 8: Vulnerability for LSOAs in the provided dataset. Notice the "border regions" that surround South and the Scotland. These all have very low (coloured black) vulnerability estimates, however this does not reflect the true vulnerability in these LSOAs. Rather, it is because SSEN service a subset of households in this LSOA, but the vulnerability measure is computed using the total number of households in the LSOA.

This artefact in the data could have a significant impact on what patterns and relationships the model learns from the data. To ensure the model learns true relationships between LSOA demographic features and vulnerability (and doesn't learn random errors in the data, which would not be useful to SSEN), we first identify these border LSOAs and withhold these from training. This way, the model is calibrated to learn relationships from only the LSOAs where the vulnerability measure is reflective of the true vulnerability in the LSOA. Importantly, the trained model can still make predictions and generate explanations for the border LSOAs, since the vulnerability metrics are not used here – recall, the model *predicts* vulnerability using other demographic data. Therefore, we are still able to generate explanations and understand the drivers of vulnerability for these border LSOAs, ensuring no customers are left behind in our analysis. When we discuss the drivers of vulnerability in LSOAs in section 2.5, we do so for ***all*** LSOAs.

2.3.5 Model performance

Having performed feature selection and identified candidate models to apply to the data, we now assess which model best captures the relationships present in the data. To determine this, we perform hyperparameter tuning for a set of candidate models, identifying the optimal learning

parameters for each. For each model we perform cross validation as described in section 2.3.2, assessing the models' performance by the prediction error each given model obtained when applied to the withheld validation data partition. We show results for this in Figure 9.

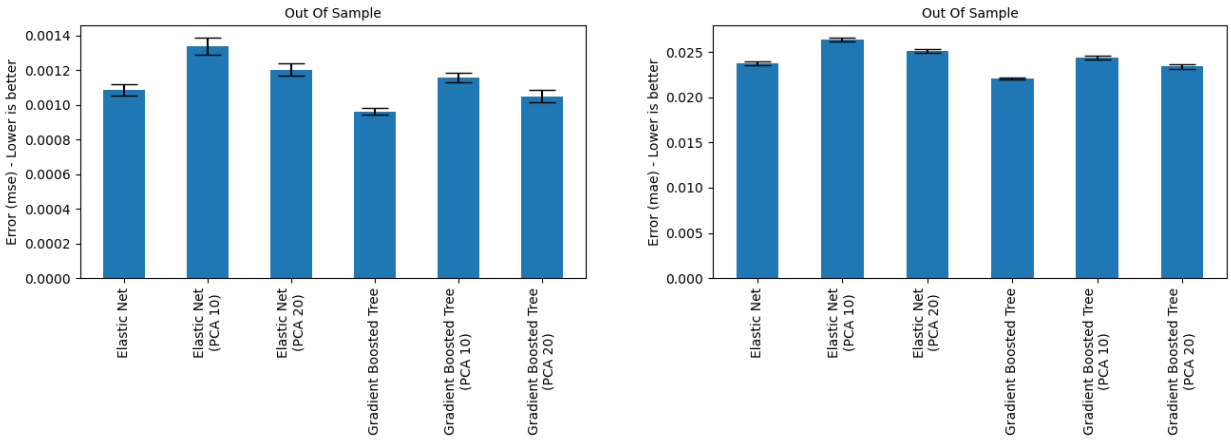


Figure 9: Comparison of candidate models, considering the mean square error (mse) and mean absolute error (mae) through 5-fold cross validation. The optimal model is the one that achieves lowest error, which we see is the Gradient boosted tree in both cases.

From Figure 9, we can question what model is optimal by finding the one that achieves the lowest error. These errors are computed through cross-validation, so the bars represent the average error across 5 different subsets of the data. The variation across these 5 subsets is shown with the error bars and is generally negligible. Inspecting the results from both error metrics, we see the optimal model is a gradient boosted tree, achieving lower error than the other candidate models for both metrics considered.

In addition to comparing a gradient boosted tree to an elastic net, we also considered if applying dimensionality reduction to the data improved either model. One way to do this is through a technique known as principal component analysis (PCA), which attempts to explain the variability present in the data using a smaller number of 'principal components' that replace the original features. While this can lead to performance improvements in some applications, we see in Figure 9 that it is not the case for this dataset, as all models with PCA have a higher error than the same type of model without PCA.

In light of these results, the final model applied in this project is a gradient boosted tree, taking in the LSOA demographic features and outputting a prediction of the fraction of households in the input LSOA that are vulnerable.

2.4 Generating explanations

To determine how the prediction model arrives at its prediction for a given LSOA, we use a method known as Shapley Additive Explanation Values (SHAP)[2]. This is the current state-of-the-art

[2] Highly efficient methodologies (Tree-SHAP) exist for computing SHAP values for tree-based models, that reduce computation time by several orders of magnitude compared to model-agnostic methods, making a gradient boosted tree model an attractive choice for this approach.

methodology in the field of machine learning explainability and is widely used when insight into complex models is required. SHAP takes a prediction model and an input (LSOA) we wish to generate explanations for and determines how each feature contributes to the final prediction the model makes. The prediction the model generates for the LSOA is broken down into a 'baseline' value and the contribution from each feature the model uses. Here, the baseline represents a starting point for comparisons, and represents the average model output across all LSOAs. This then means features increase or decrease vulnerability prediction relative to this baseline. An example output from applying this methodology on a single LSOA is shown in Figure 10.

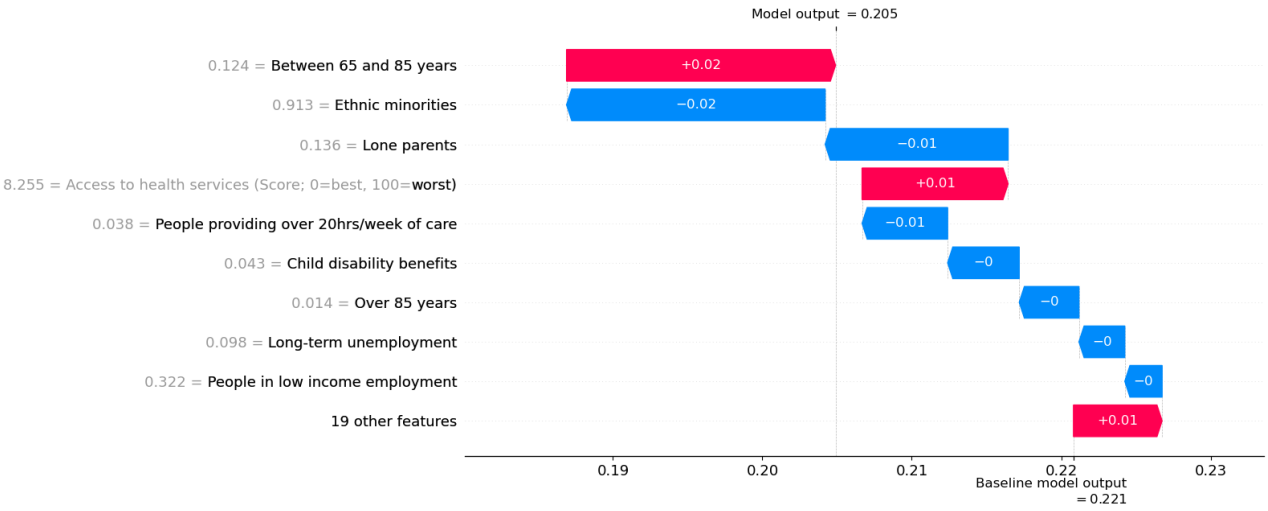


Figure 10: Explanations generated for a single LSOA, detailing how the model arrived at the prediction for this specific LSOA. Starting from a baseline value of 0.221, the model output was increased by the features shown in red (for example 'Between 65 and 85 years' and 'Access to health services score') and decreased by the features shown in blue (for example, 'Ethnic minorities' and 'Lone parents'). The given LSOA is in Brent with ID E01000471. Note that values are rounded, so may display as 0 but might contribute a small but non-zero amount.

In Figure 10, we show results for a single LSOA in Brent, with ID E01000471. We break down the contributions that each feature made to the model's final prediction. Across the entire dataset, the baseline model output is 0.221. For this specific LSOA, the model output was 0.205, meaning lower vulnerability than the baseline. We can see how each LSOA demographic feature contributed to this difference. Features that increased the model output are shown in red, in this example 'Between 65 and 85 years' and 'Access to health services score'. Features that decrease the model output are shown in blue, in this example 'Ethnic minorities' and 'Lone parents'. The numerical contributions shown here are the SHAP values for each feature. We repeat this analysis for every LSOA in the dataset to understand how each LSOA's demographic features influence its predictions.

Using this methodology, we can also assess what the most important features to the trained model are across all LSOAs. These results are shown in Figure 11, providing a global view of the analysis

compared to an LSOA-specific view given in Figure 10.

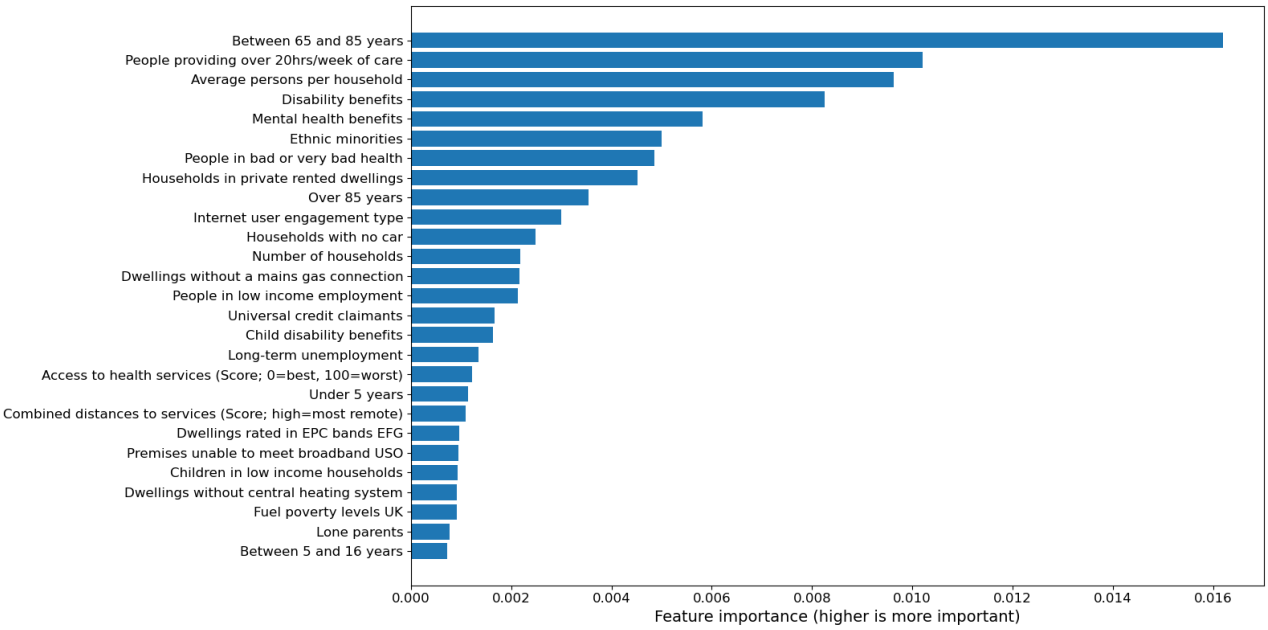


Figure 11: Global feature importance of the model. Features that are more important to the model have a larger bar in this plot, meaning the most important features are 'Between 65 and 85 years' and 'People providing over 20hrs/week of care'. The least important features by comparison are 'Between 5 and 16 years' and 'Lone parents'.

The model identifies old age, care, disability and mental health benefits as the four most important features. This is consistent with SSEN's expectations based on domain knowledge of energy vulnerability.

2.5 Identifying LSOAs with similar drivers of vulnerability

2.5.1 Identifying groupings

Having generated explanations and identified the demographic factors driving vulnerability in each LSOA, we now wish to obtain groupings of the LSOAs, such that LSOAs in the same group have similar drivers of vulnerability. In mathematical terms, this constitutes clustering the ***explanations*** for each LSOA such that LSOAs with similar explanations are placed in the same cluster (or group). There are several ways to do this, however, first note that before identifying groups we have removed explanations for the 'Scotland_YN' feature. This feature was added by us to remove any confounding effects of country-wide differences not captured by SSEN's demographic data. We exclude this feature from our analysis of the subsequent ***explanations***.

Recall that our model has 28 different features that it can use to identify vulnerability (see Table 1). This is a high dimensional space and can produce poor clustering results. To remedy this, we first apply dimensionality reduction to the explanations, mapping them from a 28-dimensional space to a 5-dimensional space. This is done through a methodology known as 'Uniform Manifold Approximation and Projection' (UMAP). This is a widely used and state-of-the-art methodology for data reduction of this type and has been successfully used to cluster SHAP values in other applications. 5-dimensions was chosen to balance the amount of noise introduced as dimensionality increased, with the oversimplification of the data that is possible as dimensionality

decreases. Having reduced the dimensionality of the clustering problem, we then apply a range of clustering methods to assess how well the data clusters, and how many clusters exist in the data. These were k-means clustering, hierarchical agglomerative clustering (HAC), and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). This offers distinct mathematical approaches to the problem that ensure different types of patterns can be explored in the data. HDBSCAN shows little success on this data, so we do not proceed with this method.

2.5.2 Clustering methods

Two common clustering methods in the field of machine learning are K-means and HAC. Here we detail how each work, to give an intuitive understanding rather than a rigorous mathematical understanding.

K-means is perhaps the most well-known clustering method, in which the modeller selects the number of clusters in their data (denoted “k”) and the data is then searched to best segment it into these k groups. This is done by randomly initialising k cluster centres and assigning each data-point into the cluster whose centre it lies closest to. Having assigned all the data-points, the cluster centres are then re-computed as the average location of all data-points that lie in that cluster. Cluster assignments are then recomputed using these new cluster centres. This process is repeated until the algorithm converges to a stable solution and the cluster assignments and centres no longer change.

HAC is a fundamentally different approach to clustering. Whilst k-means split the data into k groups and then refined these groups until there were no further changes, HAC starts by assuming each data-point is a unique cluster. We then ask, of all possible pairs of data-points, which two are most similar? These then get merged into a cluster, and the question is asked again. This is repeated over and over, merging data-points either with other data-points, or into other clusters depending on what they are most similar to in the data. After a sufficient number of merges, the entire dataset is then contained in a single cluster. One visualises how datapoints were merged by looking at this sequence of actions and constructing a dendrogram (forward reference to Figure 12 for an example in this context). A dendrogram shows each individual data-point at its base, and as one moves up the dendrogram, we see the clusters emerge. The final clustering is achieved by selecting a height at which to “cut” this dendrogram, that is a height on the y-axis that corresponds to a particular clustering of the data. In the context of a dendrogram, a “height” or “distance” between points or clusters is determined through the choice of linkage function. A linkage function simply tells us how similar two clusters are, with different choices of function altering exactly how similarity is defined. For this work, we use the ‘Ward’ linkage function, which is a widely used linkage function that is known to analyse the variation in clusters.

2.5.3 Selecting the number of clusters in the data

We first must determine how many clusters are present in the data. If we choose too low a number, then much of the detail present in the data may be lost. Equally, if we choose too high a number, then meaningful relationships may become distorted by noise in the results. To determine how many clusters there are in the data, we contrast two methods. The first is through repeated applications of k-means clustering, where the number of clusters specified is incrementally increased. For each result, we then inspect each LSOA and the cluster it falls into. We compute

the distance between each point and the centre of its associated cluster and repeat this for all LSOAs. This is a measurement known as inertia and is shown in Figure 12 (left). For low numbers of clusters, the inertia will be very high, as there are very few cluster centres and hence points in each cluster are far from these centres. However, as the number of clusters becomes large, the points in each cluster will naturally fall closer and closer to their centres. The changes in distance will diminish at some point, suggesting the data has been sufficiently split to explain its variation in space.

A second approach for selecting the number of clusters utilises HAC methods, specifically building a tree structure known as a dendrogram, explained previously in section 2.5.2. Visual inspection of the resulting dendrogram can suggest natural numbers of clusters in the data. Such a plot is shown in Figure 12 (right).

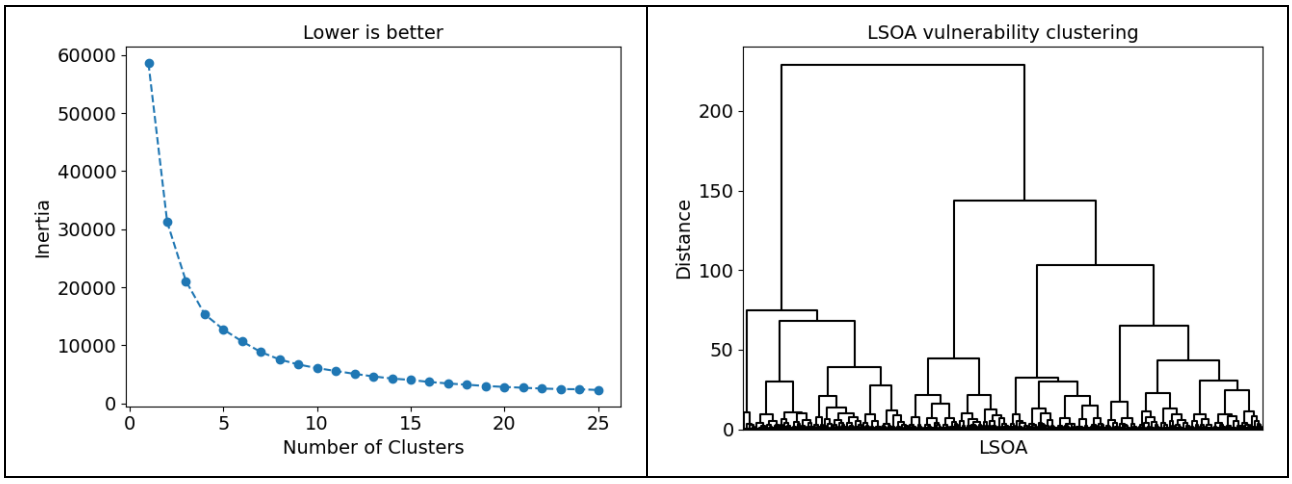


Figure 12: Inertia plot (left) and dendrogram (right) assessing how many clusters might exist in the data. There is a significant depreciation in the change in inertia between 5-10 clusters, and the dendrogram suggests around 7 clusters are preened before significant noise appears.

Inspecting Figure 12, we see similar results from both the inertia plot (left) and dendrogram (right). From the inertia plot, there is a vast decrease in the inertia as the number of clusters increase from one to three. This is expected, and often observed in clustering across a range of applications. However, the decrease in inertia becomes less prominent after this, and after fifteen clusters there is almost no change in inertia. The plot suggests that between five and ten clusters will capture most of the meaningful variation in the data.

Inspecting the dendrogram Figure 12 (right), we have a different approach to clustering the data. We can see that at distances below around twenty-five, thousands of LSOAs are merged into pairs or small groups. Recall that there are 6009 LSOAs in this dataset. As the distance increases above this, there are fewer and fewer groupings. If one cuts this dendrogram at a distance of fifty, they attain 7 clusters. As one cuts the dendrogram lower than this, small changes in distance start to rapidly change the number of clusters one attains. This indicates too fine a structure past this point.

From this analysis, a sensible and logical choice for the number of clusters in the data is 7, and this is used from this point on.

3 Results

From our analysis in section 2.5, applying k-means clustering to group the model explanations into 7 clusters is supported by the data, and there is no clear benefit to using the alternative clustering techniques explored. As a result, we proceed with this and now move on to understanding the main drivers of vulnerability that are common to all LSOAs in each given cluster. A first step towards this is to compare the average vulnerability levels of each cluster, measured by the sum of the SHAP values. This is given in Figure 13.

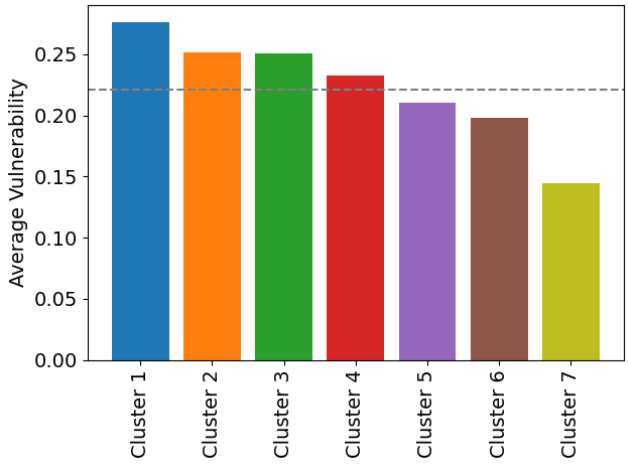
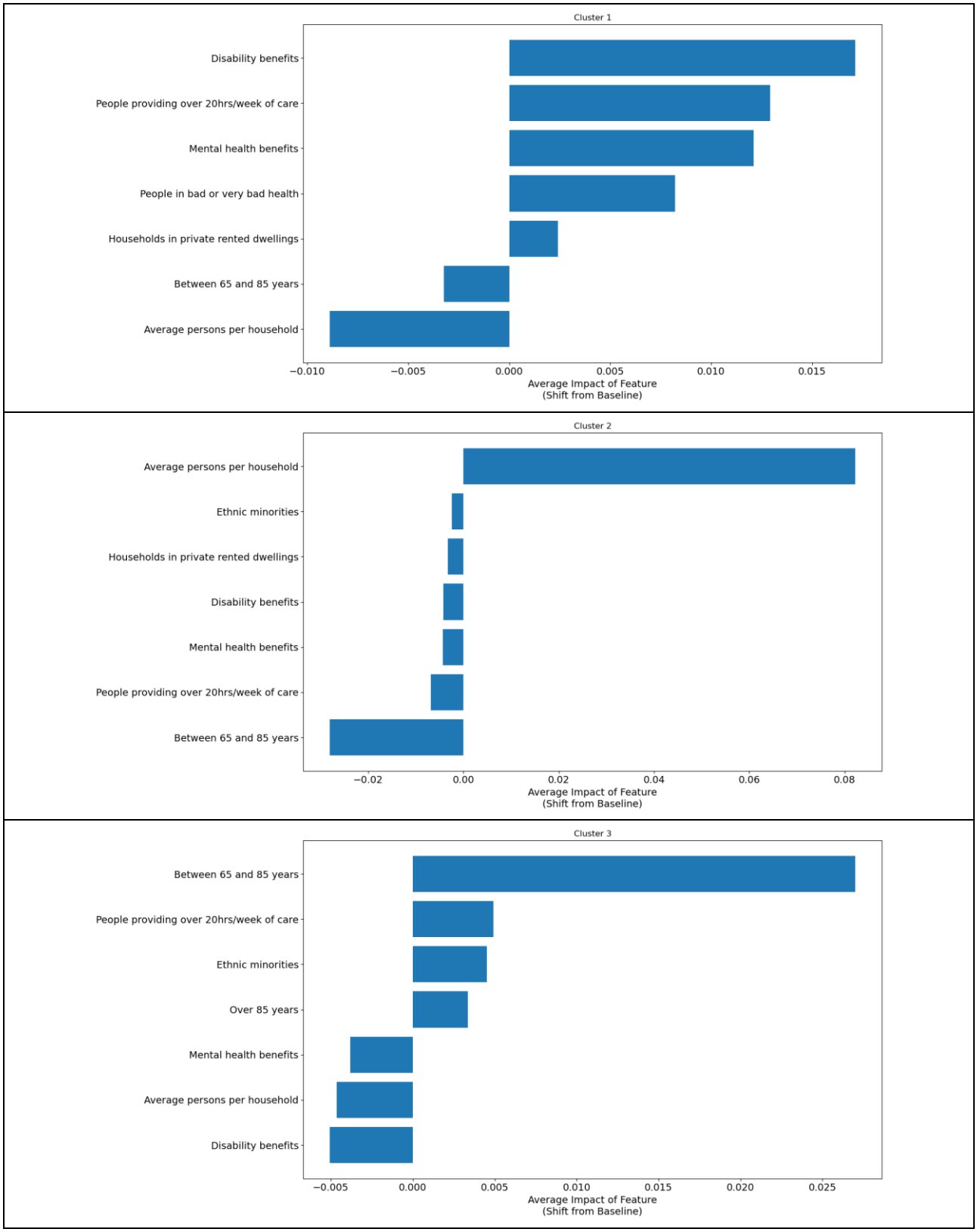
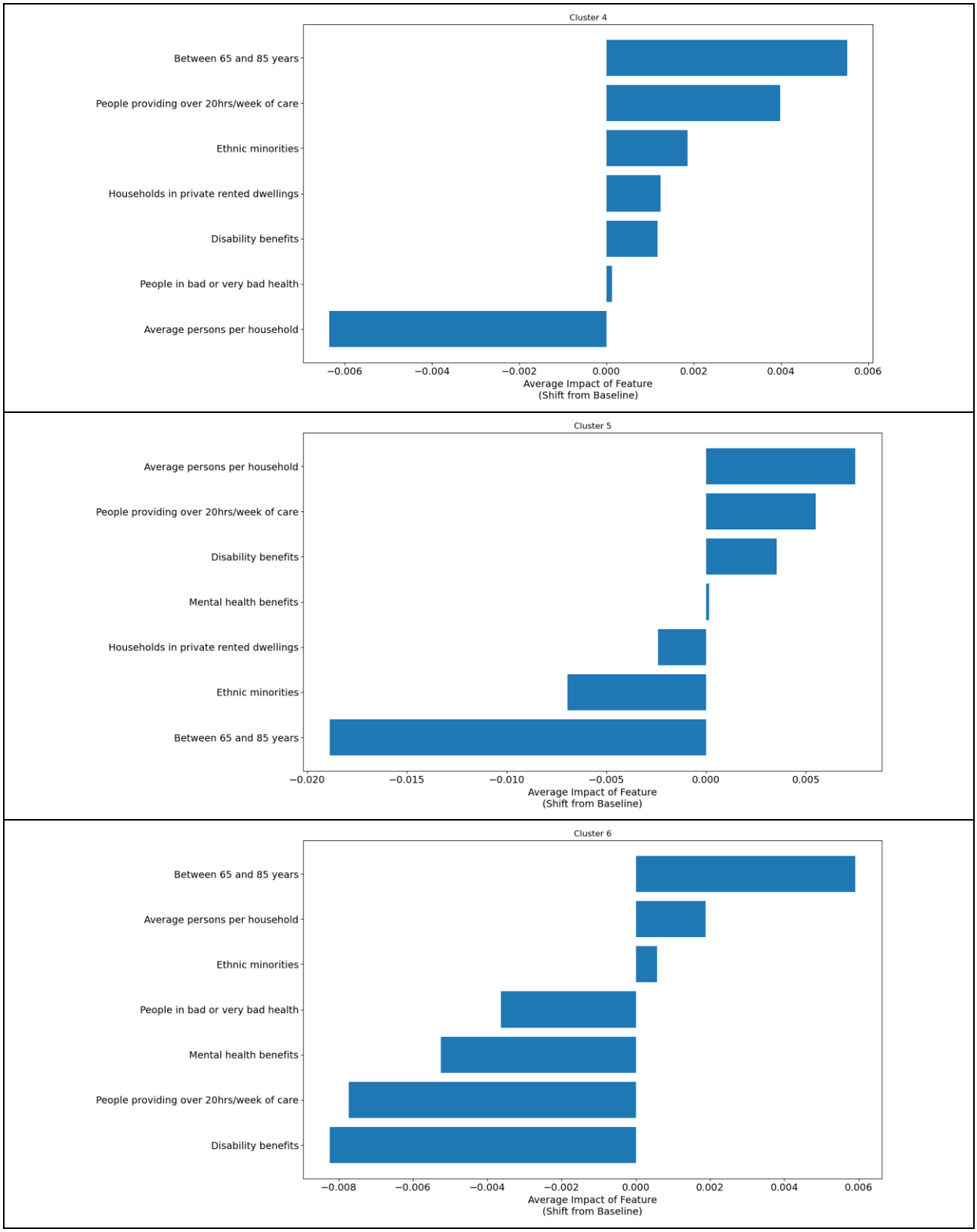


Figure 13: Average vulnerability for the LSOAs in each cluster, measured by summing the SHAP values for all model features other than Scotland_YN as discussed. Clusters are ordered such that the cluster with highest average vulnerability is cluster 1, and the cluster number increases as average vulnerability decreases. The average vulnerability across all LSOAs is shown with a dashed grey line for reference.

The figure shows that cluster 1 has substantially higher average vulnerability than any cluster, while on the contrary, cluster 7 has a much lower average vulnerability relative to all other clusters. However, we are not just interested in the overall levels of vulnerability within each group, but also ***identifying the drivers of vulnerability in each cluster.*** To do this, we can inspect the most influential demographic characteristics within each cluster – measured by considering the features that had the highest SHAP values averaged across all LSOAs within the cluster. We show the 7 most significant drivers in each cluster in Figure 14. Note that the x-scales in this plot differ for each cluster.


Cluster 1
Disability benefits
People providing over 20hrs/week of care
Mental health benefits
People in bad or very bad health
Households in private rented dwellings
Between 65 and 85 years
Average persons per household
−0.010
−0.005
0.000
0.005
0.010
0.015
Average Impact of Feature
(Shift from Baseline)
Cluster 2
Average persons per household
Ethnic minorities
Households in private rented dwellings
Disability benefits
Mental health benefits
People providing over 20hrs/week of care
Between 65 and 85 years
−0.02
0.00
0.02
0.04
0.06
0.08
Average Impact of Feature
(Shift from Baseline)
Cluster 3
Between 65 and 85 years
People providing over 20hrs/week of care
Ethnic minorities
Over 85 years
Mental health benefits
Average persons per household
Disability benefits
−0.005
0.000
0.005
0.010
0.015
0.020
0.025
Average Impact of Feature
(Shift from Baseline)


Cluster 4
Between 65 and 85 years
People providing over 20hrs/week of care
Ethnic minorities
Households in private rented dwellings
Disability benefits
People in bad or very bad health
Average persons per household
−0.006
−0.004
−0.002
0.000
0.002
0.004
0.006
Average Impact of Feature
(Shift from Baseline)
Cluster 5
Average persons per household
People providing over 20hrs/week of care
Disability benefits
Mental health benefits
Households in private rented dwellings
Ethnic minorities
Between 65 and 85 years
−0.020
−0.015
−0.010
−0.005
0.000
0.005
Average Impact of Feature
(Shift from Baseline)
Cluster 6
Between 65 and 85 years
Average persons per household
Ethnic minorities
People in bad or very bad health
Mental health benefits
People providing over 20hrs/week of care
Disability benefits
−0.008
−0.006
−0.004
−0.002
0.000
0.002
0.004
0.006
Average Impact of Feature
(Shift from Baseline)

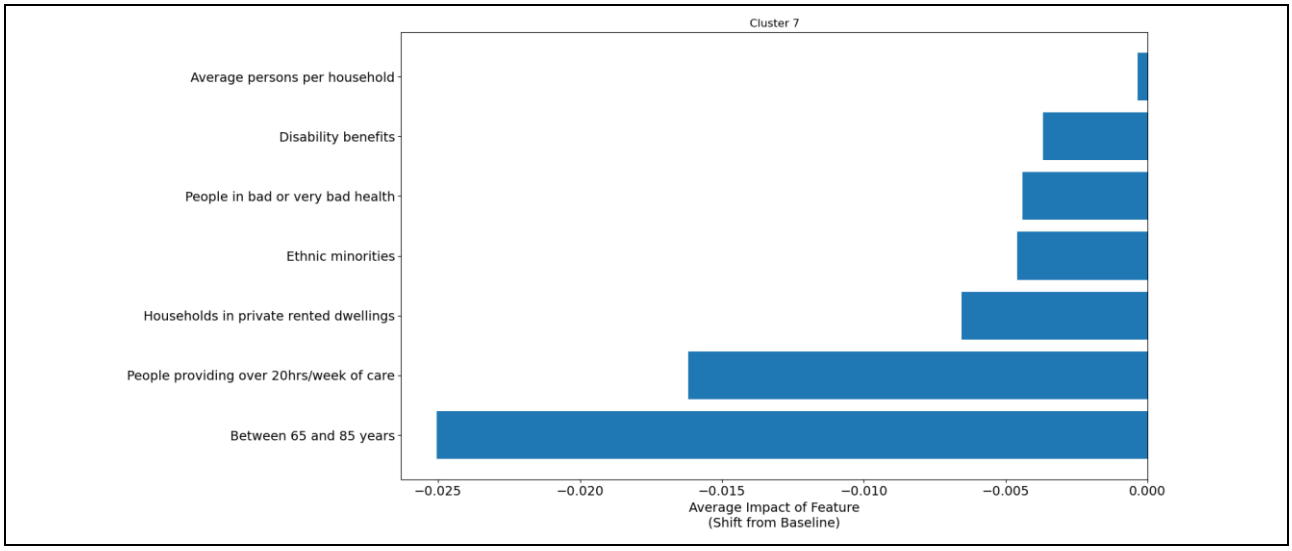


Figure 14: Visualisation of the 7 most influential features in each cluster. These are the 7 features that have the highest average absolute SHAP value. There are both similarities across clusters but also differences. Age, health and benefits are often strong drivers in vulnerability, either increasing or decreasing it, but other features only influence a single cluster, or do so far more strongly than in any others. Note that the x-scales on these plots differ for each cluster.

From Figure 14, there are both similarities across clusters but also differences. Age, health and benefits are often strong drivers in vulnerability, either increasing or decreasing it, but other features can be seen to be far more prominent in one cluster than any of the others. To complement Figure 13, in Table 2 we provide qualitative summaries of the most influential demographic features within each cluster.

Cluster Number	Description of cluster
1	Driven up by higher levels of poor health and disability/mental health benefit claimants, reduced by smaller household sizes.
2	Driven up by larger household sizes, reduced by lower elderly population levels.
3	Driven up by larger elderly population levels, reduced by lower levels of disability and mental health benefit claimants.
4	Driven up by larger elder population levels and moderately higher provision of care, reduced by smaller household sizes.
5	Driven down by lower elderly population levels and larger levels of ethnic diversity, increased by higher household sizes and greater provision of care.
6	Driven down by lower level of bad health and disability/mental health benefit claimants, increased by moderate elderly population levels and household sizes.
7	Driven down by substantially lower elderly population levels, less provision of care and a higher level of households in private rented dwellings.

Table 2: Interpretation definitions of clusters – detailing what drives vulnerability in each of the 7 identified clusters.

Interesting results arise from inspection of the clustering results. The feature ‘households in private rented dwellings’ strongly influences vulnerability in cluster 7 but is not observed as a main driver in any of the other clusters. Many other demographic features arise in multiple clusters, for example old age, provision of care and household sizes, however the average vulnerability in each

cluster differs. We visualise the distribution of relevant demographic features in the appendix, to offer some insight into how these demographic features compare across clusters.

A further analysis one might ask is, how many LSOAs and households fall into each cluster? This gives a gauge of how dominant these groups of drivers of vulnerability are across the LSOAs that SSEN service. This is shown in Figure 15.

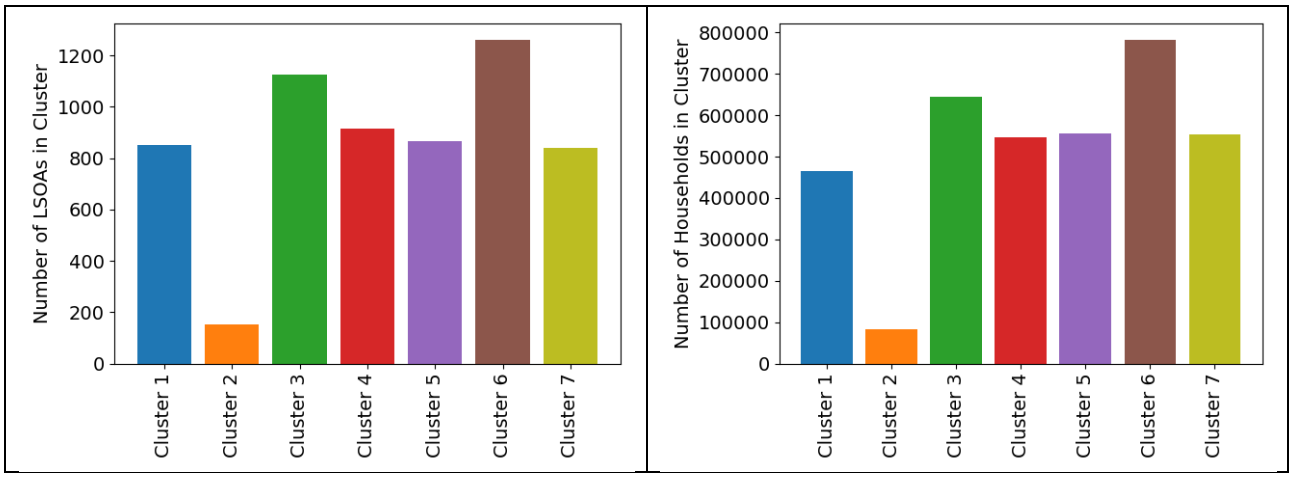


Figure 15: Cluster sizes, detailing how many LSOAs fall into the discovered clusters (left) and how many households there are in each cluster (right). Cluster 2 is by far the smallest, encompassing less than 200 LSOAs, whereas cluster 6 is the largest, encompassing over 1200.

From Figure 15, we see that cluster 1, showing very high vulnerability and driven mainly by disability and mental health benefit claims, poor health, and provision of care encompasses slightly over 800 LSOAs. Cluster 7, showing very low vulnerability typically driven by levels of old age, provision of care, households in private rented dwellings and levels of ethnic minorities encompasses around the same number of LSOAs. Clusters 1, 4, 5 and 7 are reasonably similar in size (the number of LSOAs they encompass) however cluster 2 is by far the smallest, meaning that under 200 LSOAs have high vulnerability with increases typically driven by household sizes and decreases typically driven by elderly population levels. Whilst this cluster is small, it is useful to see it in the data, as it ensures that when making investments in the future, these smaller but present groups of LSOAs can be accounted for. The two largest clusters are 3 and 6.

Finally, we visualise where these clusters fall geographically, to provide insights into whether the levels and drivers of vulnerability identified by our groupings can be associated with specific regions of Great Britain. This is shown in Figure 16.

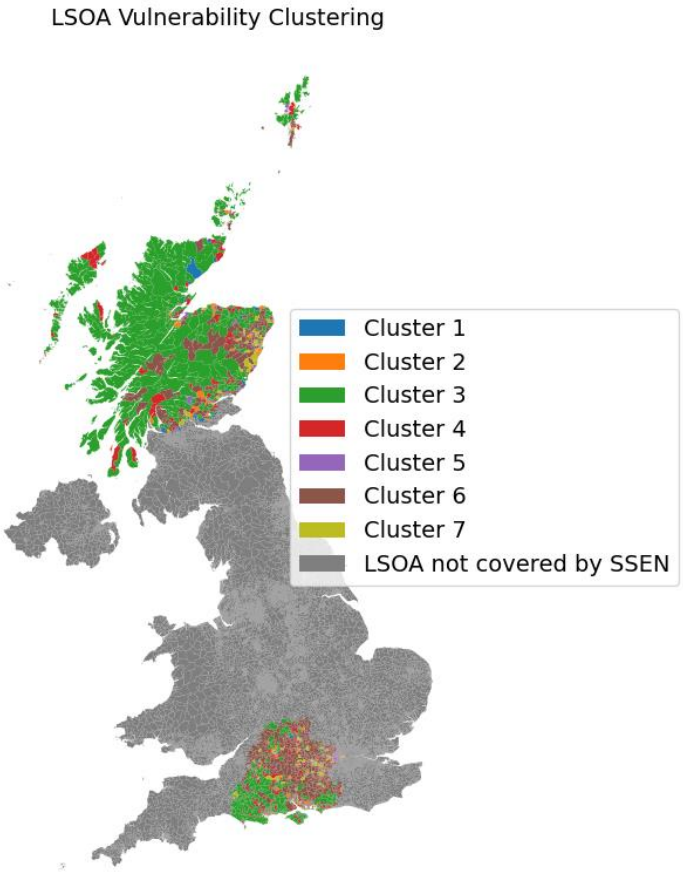


Figure 16: Map of the UK, with LSOAs coloured by the cluster they are a member of. LSOAs that are not serviced by SSEN are coloured in grey. Note that clusters are generally not solely in Scotland or solely in the south, they cover LSOAs in both. Large parts of Scotland belong to cluster 3, whereas large parts of London belong to cluster 6.

Inspecting Figure 16, we can see that clusters are generally not solely in Scotland or solely in the south, they cover LSOAs in both. However, there is still clear spatial structure present. Large parts of Scotland and the Southwest of England belong to cluster 3 (recall, high vulnerability typically increased by the levels of old age and provision of care, with vulnerability typically reduced by low levels of disability and mental health benefit claims). Therefore, investment in these locations may be best placed in addressing these drivers. Large parts of London on the other hand are split between clusters 3 (recall, high vulnerability typically increased by the levels of old age and provision of care, with vulnerability levels typically reduced by low levels of disability and mental health benefit claims) and 6 (low vulnerability typically increased by old age and household sizes, with vulnerability typically decreased by levels of disability and mental health claims, as well as provision of care and levels of bad health). This alludes to not only different properties of areas to address within London, but also different demographic factors to target with investments. The relationships between LSOAs spatially and their drivers of vulnerability suggest that specific investments may benefit Scotland more so than the South of England. For example, investments that address the underlying drivers of vulnerability in cluster 3 will impact large parts of Scotland and the Southwest of England, but less so central London. Clearly, future investment relating to

vulnerability can be targeted to address the different drivers, and different locations in Great Britain will benefit from different investments.

As described in section 1.1.1, the results in this section consider vulnerability computed using the actual PSR data. In the appendix, we repeat the analysis, but compute vulnerability using the estimated PSR eligibility and comment on the differences observed.

4 Conclusion

In this project, we have performed a mathematical and data-driven analysis of vulnerability across the LSOAs that SSEN service. For each LSOA, we have determined how each demographic feature contributes to that LSOA's vulnerability. Finally, we have used this to determine groups of LSOAs that share similar drivers of vulnerability. Using these groupings, we have gained an understanding of what fundamentally drives vulnerability across Scotland and the south of England. Using this, SSEN can make network investment decisions that account for consumer vulnerability going forward, and ensure no consumers are left behind during the rapid transitions taking place in the energy sector. It also aids SSEN in meeting the sustainability development goals, particularly ensuring clean and affordable energy to all, reducing inequality, and working towards sustainable cities and communities.

We found that 7 distinct groups of LSOAs exist in the data, where each group has different drivers of vulnerability. For each of these 7 groups, we isolated what the primary drivers of vulnerability are, seeing that age, provision of care, disability and mental health benefits and household sizes are common drivers of vulnerability in many LSOAs. However, we observed that these demographic factors impact different LSOAs in different ways. Additionally, there are specific demographic factors that drive vulnerability for one group of LSOAs but have little impact on the other groups. A clear example of this is the households in private rented dwellings, which strongly influences vulnerability in only one of the clusters of LSOAS we have discovered.

We discovered three groups of LSOAs that have particularly high vulnerability, and hence warrant particular attention when making network investment decisions: these groups of LSOAs are areas where customers would be particularly disadvantaged by the risk of frequent power cuts, and would benefit from a more resilient future network. Our analysis shows vulnerability in the first of these groups is increased by levels of disability and mental health benefits, poor health and provision of care. Vulnerability in the second group is typically increased by household sizes. The third group shows that old age and provision of care increase vulnerability. In addition to this, we detail what decreases vulnerability in each of these clusters. These three areas warrant specific network investments to ensure no customers are left behind in the ever-changing energy landscape.

There are several recommendations for future work, and considerations if this analysis is to be applied on a wider scale. First, note that the analysis done in this project applies to the LSOAs served by SSEN. It may be that different LSOAs, in other regions of Great Britain, show some different drivers of vulnerability: some demographic features may be more important, others less so. Care must be taken in extrapolating our results onto other LSOAs. While the results reflect the

behaviours observed in LSOAs serviced by SSEN, they may not necessarily reflect the behaviours observed in LSOAs serviced by different DNOs. To understand whether these relationships do change, the same demographic data would need to be obtained from each DNO, and the model retrained, and explanations regenerated across the Great Britain. It may also be fruitful to correct for locational differences, as has been done between Scotland and the South of England in this project. We recommend the same form of analysis is conducted on all LSOAs across Great Britain, to ensure vulnerable customers are considered in investment decisions for all DNOs.

Further, we are aware that SSEN wish to understand not only vulnerability today, but also how this will evolve in the future. The analysis presented in this report can act as a base for this understanding. We currently understand, from an analytical perspective, what drives vulnerability today. Consideration should be given to how SSEN can forecast each of the drivers identified, and the associated uncertainty in these forecasts. Additionally, there may be emerging aspects that begin to influence vulnerability in the future, but do not influence it currently, or for which data is not currently collected. Incorporation of these factors into future investment decisions may require the unification of our data-driven methodologies with socio-economic methodologies, using these to augment each other.

We are also aware that the data provided for this study is gathered from several sources. One such source is the 2011 census, the most recent available at the time of the project, while another is the PSR data that is taken from 2022. This may mean that some demographic measurements in the data require updating, to better reflect the true status of the LSOAs and how they have changed since 2011. We recommend the same analysis pipeline is rerun when such data becomes available, and conclusions reassessed to inspect how relationships may have changed between 2011 and 2022.

Appendix

Demographic feature distributions for each identified cluster

Various demographic features have been identified that drive vulnerability in LSOAs across Scotland and the South of England. It can be useful for reference if the distribution of these features is visualised, understanding not only that a given feature is a driver of vulnerability for a particular group of LSOAs, but also what values this feature takes relative to the other LSOAs. We do so for the main set of drivers found in each of the 7 clusters in the following plots.

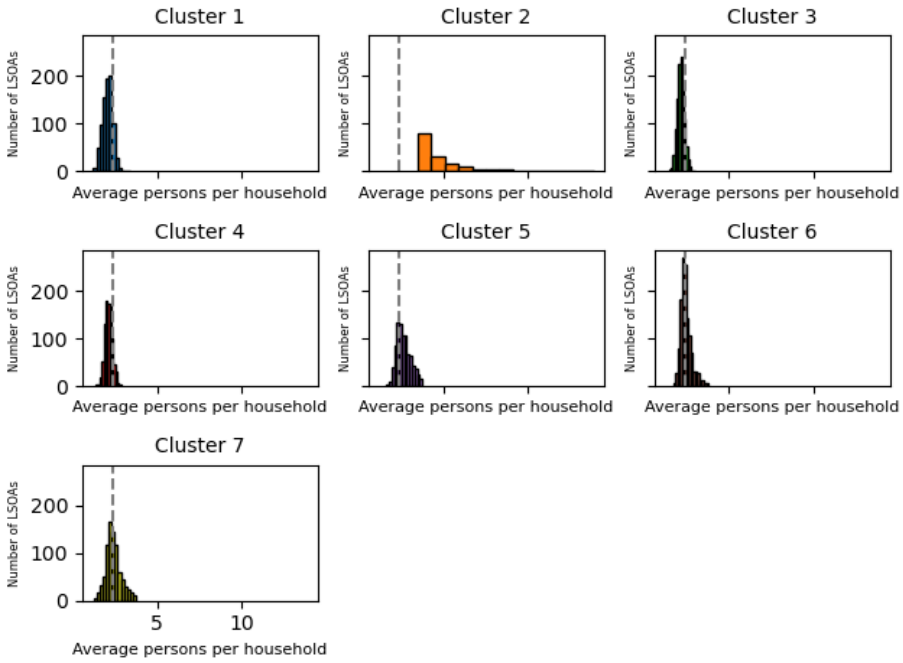


Figure 17: Distributions of the feature 'Average persons per household' for each cluster of LSOAs identified. The mean across all LSOAs is marked in dashed grey for reference.

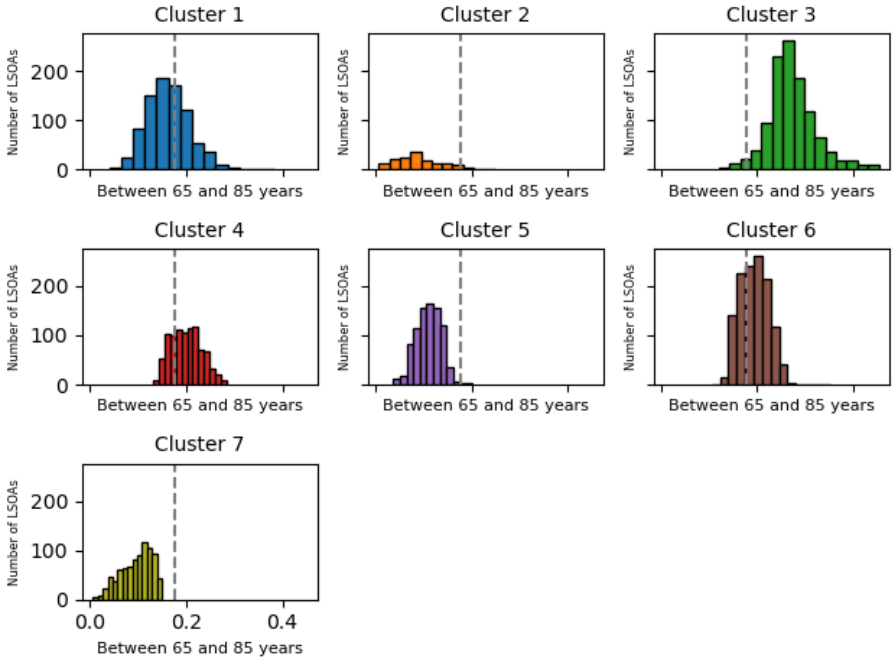


Figure 18: Distributions of the feature 'Between 65 and 85 years' for each cluster of LSOAs identified. The mean across all LSOAs is marked in dashed grey for reference.

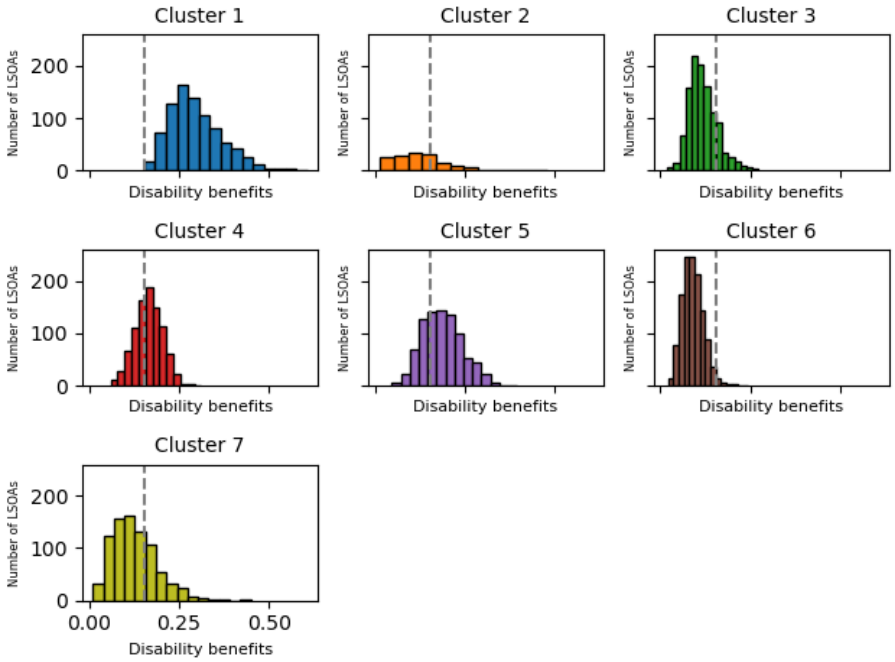


Figure 19: Distributions of the feature 'Disability benefits' for each cluster of LSOAs identified. The mean across all LSOAs is marked in dashed grey for reference.

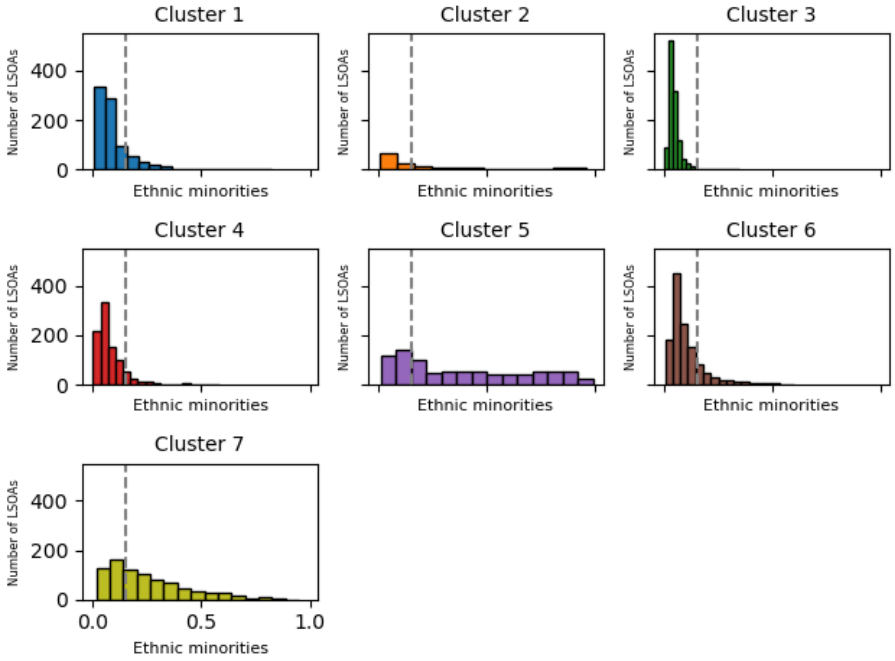


Figure 20: Distributions of the feature 'Ethnic minorities' for each cluster of LSOAs identified. The mean across all LSOAs is marked in dashed grey for reference.

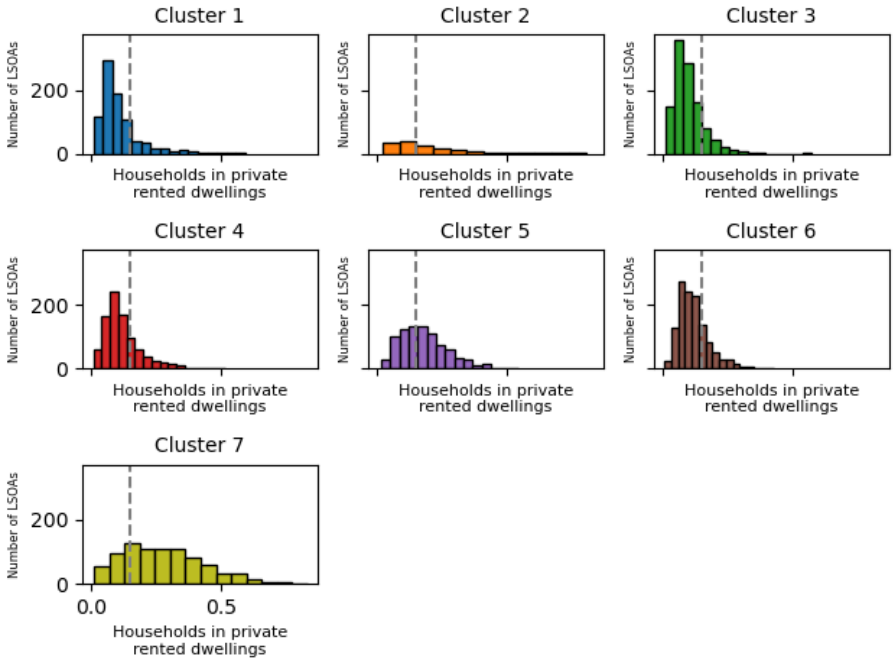


Figure 21: Distributions of the feature 'Households in private rented dwellings' for each cluster of LSOAs identified. The mean across all LSOAs is marked in dashed grey for reference.

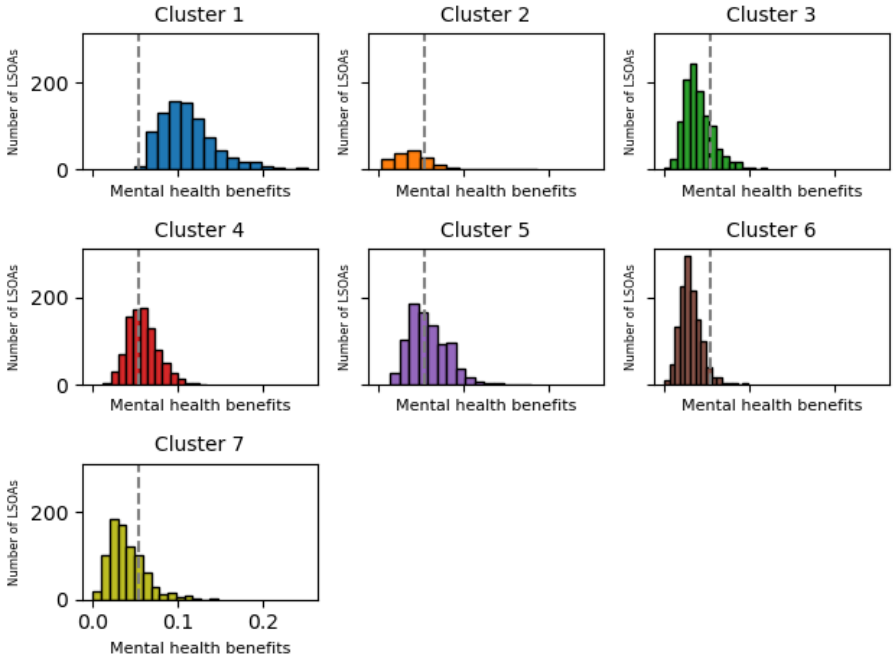


Figure 22: Distributions of the feature 'Mental health benefits' for each cluster of LSOAs identified. The mean across all LSOAs is marked in dashed grey for reference.

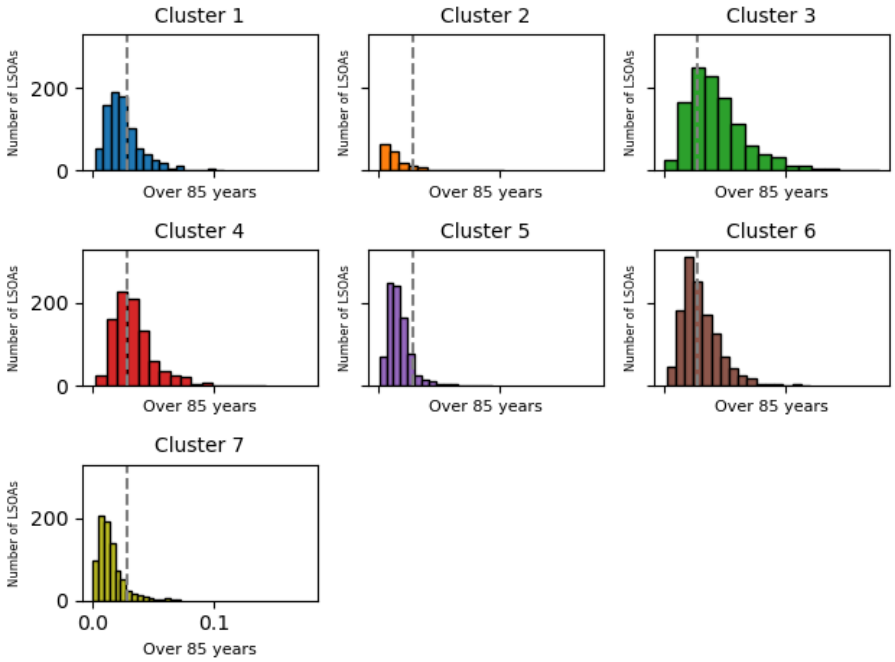


Figure 23: Distributions of the feature 'Over 85 years' for each cluster of LSOAs identified. The mean across all LSOAs is marked in dashed grey for reference.

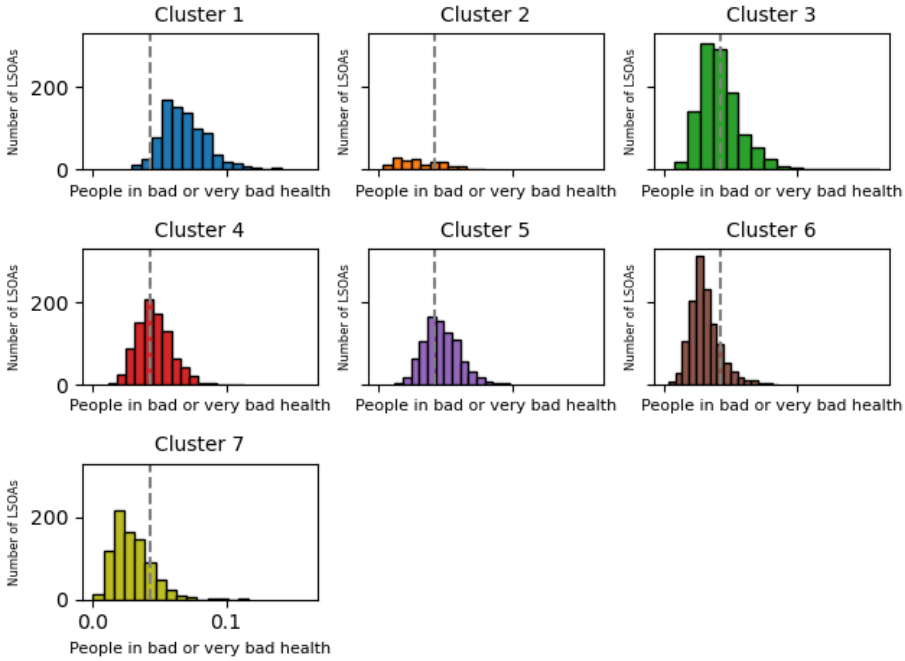


Figure 24: Distributions of the feature 'People in bad or very bad health' for each cluster of LSOAs identified. The mean across all LSOAs is marked in dashed grey for reference.

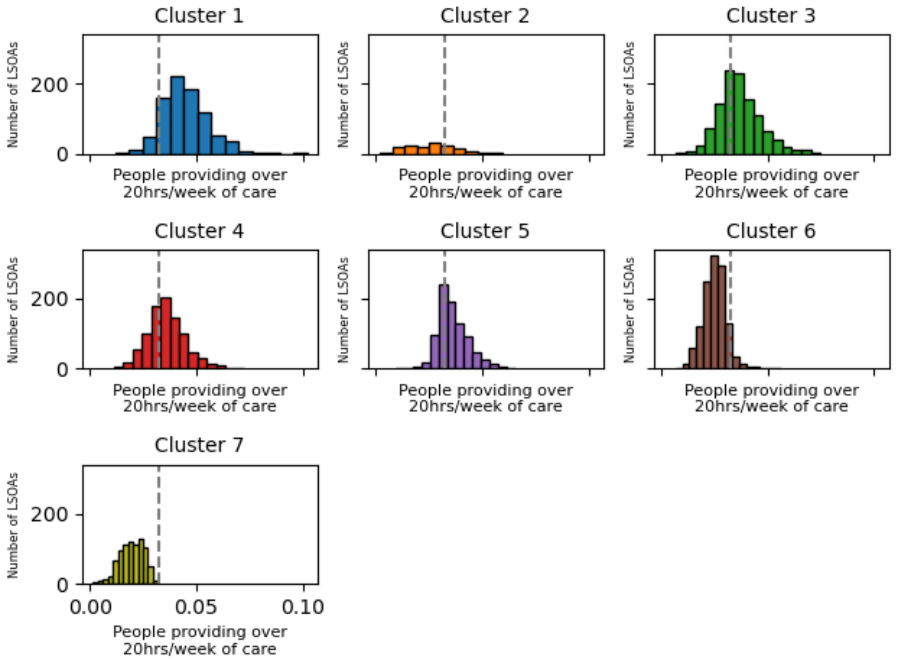


Figure 25: Distributions of the feature 'People providing over 20hrs/week of care' for each cluster of LSOAs identified. The mean across all LSOAs is marked in dashed grey for reference.

Results when considering estimated vulnerability levels

We noted in section 1.1.1 that this project considered vulnerability using the actual number of PSR records in the data. SSEN have performed extensive work to estimate the 'gap' between this and

those eligible to be on the PSR but are not for a variety of reasons. We can repeat our modelling approach using this estimated eligibility in-place of the actual PSR record data, however the results must be interpreted with care. Firstly, recall that the estimate for those eligible for the PSR depends on the demographic data we have. Therefore, in mathematical terms, a 'model' is being used by SSEN to generate this estimate, and hence whatever this criterion is will strongly influence the relationships that are uncovered in our analysis. Regardless, it can be useful to visualise these results for reference, so we include them here. We show the average vulnerability for each cluster in Figure 26, the drivers of vulnerability in each cluster in Figure 27, the cluster sizes in Figure and the spatial location of the clusters in Figure 29.

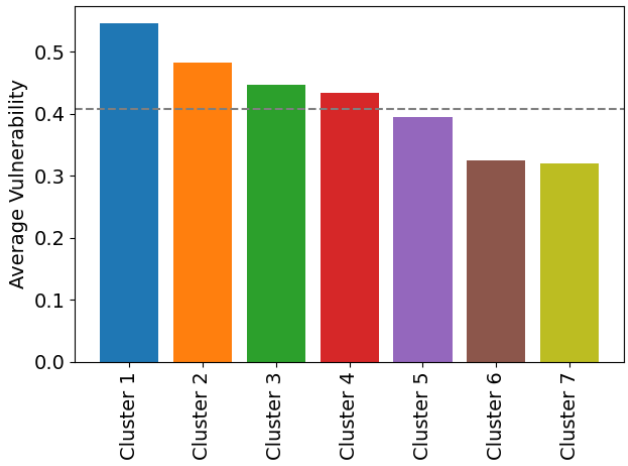
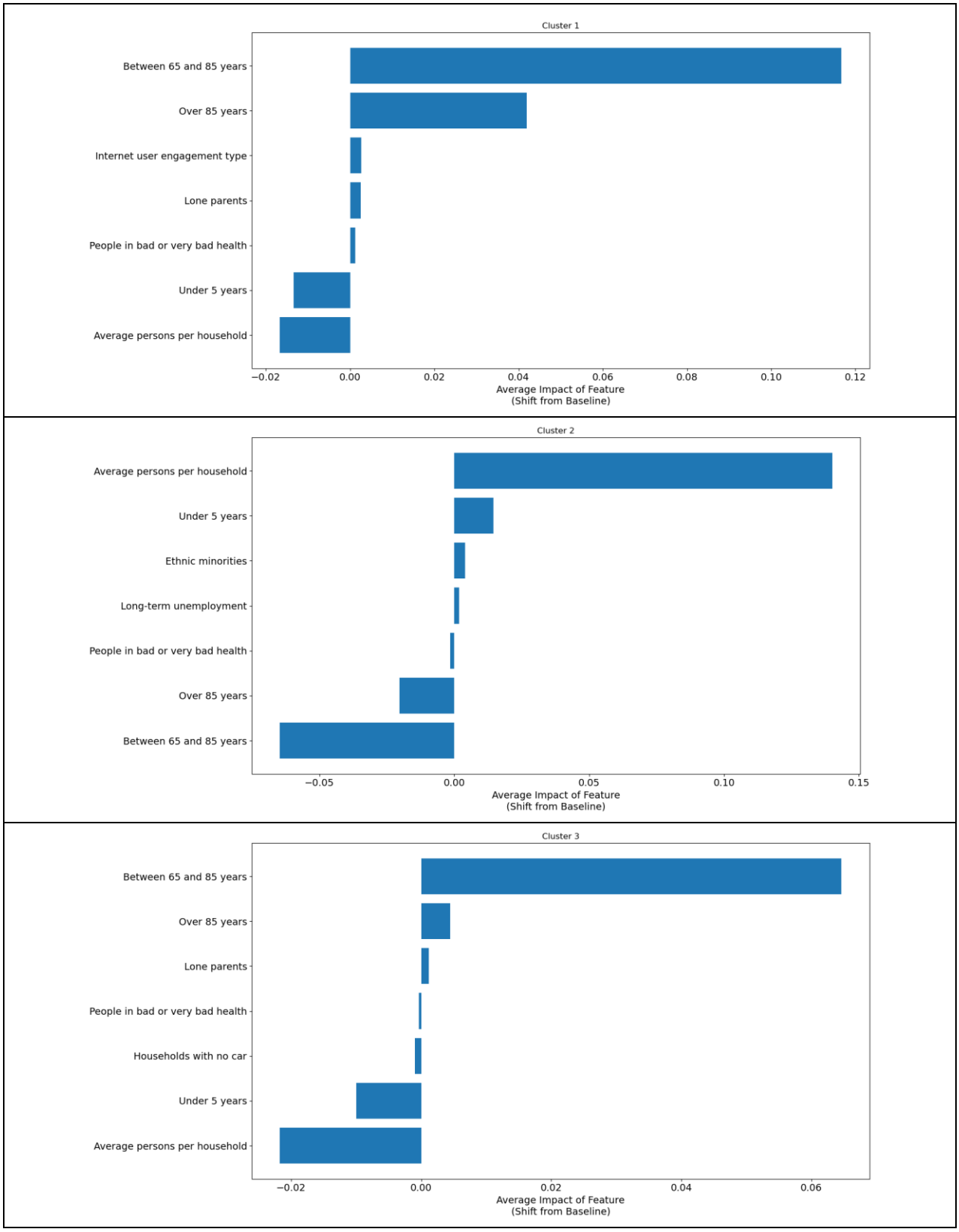
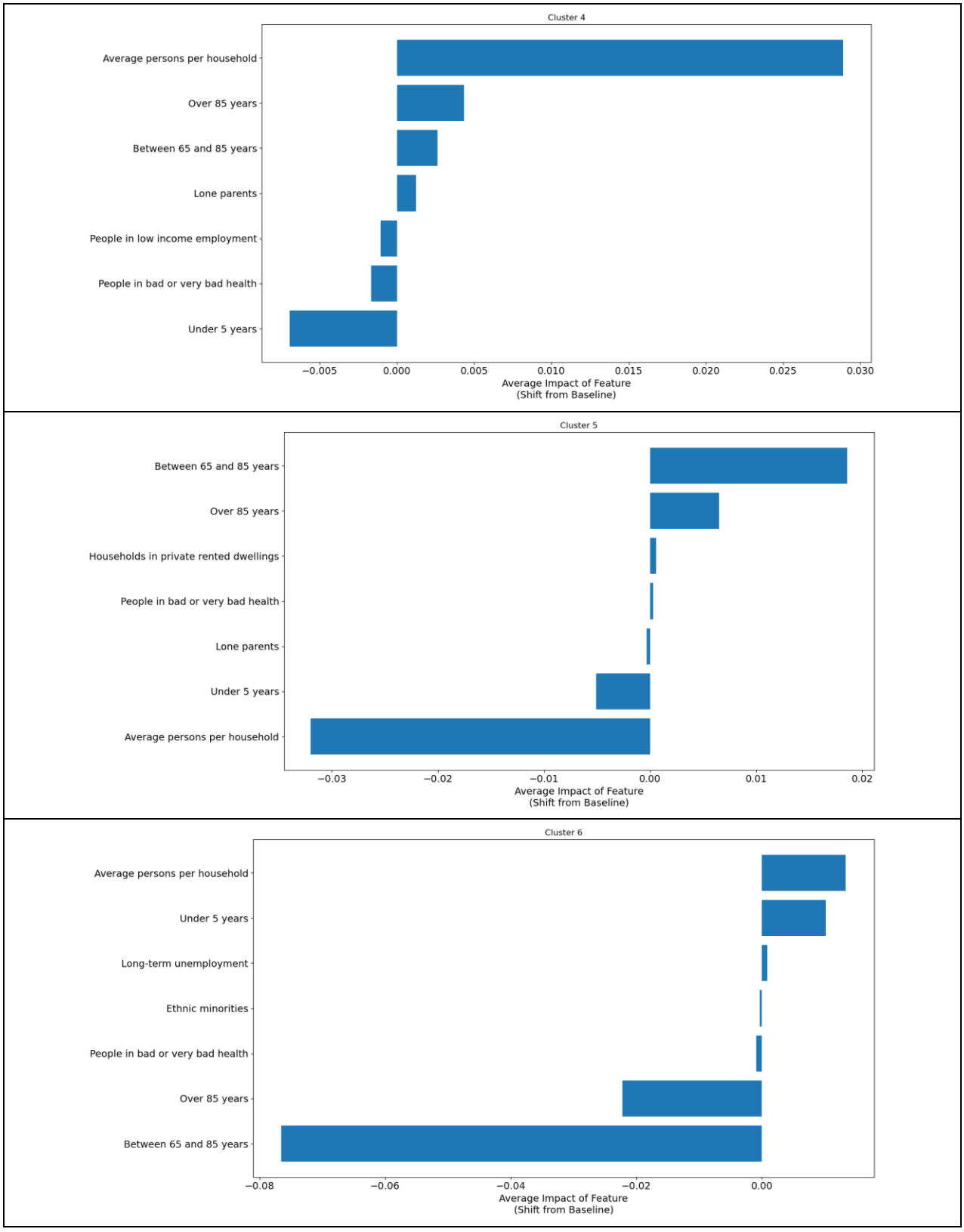


Figure 26: Average vulnerability for the LSOAs in each cluster. Clusters are ordered such that the cluster with highest average vulnerability is cluster 1, and the cluster number increases as average vulnerability decreases. The average vulnerability across all LSOAs is shown with a dashed grey line for reference. Here vulnerability is defined using the estimated PSR eligibility as computed by SSEN.


Cluster 1
Between 65 and 85 years
Over 85 years
Internet user engagement type
Lone parents
People in bad or very bad health
Under 5 years
Average persons per household
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Average Impact of Feature
(Shift from Baseline)
Cluster 2
Average persons per household
Under 5 years
Ethnic minorities
Long-term unemployment
People in bad or very bad health
Over 85 years
Between 65 and 85 years
−0.05
0.00
0.05
0.10
0.15
Average Impact of Feature
(Shift from Baseline)
Cluster 3
Between 65 and 85 years
Over 85 years
Lone parents
People in bad or very bad health
Households with no car
Under 5 years
Average persons per household
−0.02
0.00
0.02
0.04
0.06
Average Impact of Feature
(Shift from Baseline)


Cluster 4
Average persons per household
Over 85 years
Between 65 and 85 years
Lone parents
People in low income employment
People in bad or very bad health
Under 5 years
−0.005
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Average Impact of Feature
(Shift from Baseline)
Cluster 5
Between 65 and 85 years
Over 85 years
Households in private rented dwellings
People in bad or very bad health
Lone parents
Under 5 years
Average persons per household
−0.03
−0.02
−0.01
0.00
0.01
0.02
Average Impact of Feature
(Shift from Baseline)
Cluster 6
Average persons per household
Under 5 years
Long-term unemployment
Ethnic minorities
People in bad or very bad health
Over 85 years
Between 65 and 85 years
−0.08
−0.06
−0.04
−0.02
0.00
Average Impact of Feature
(Shift from Baseline)

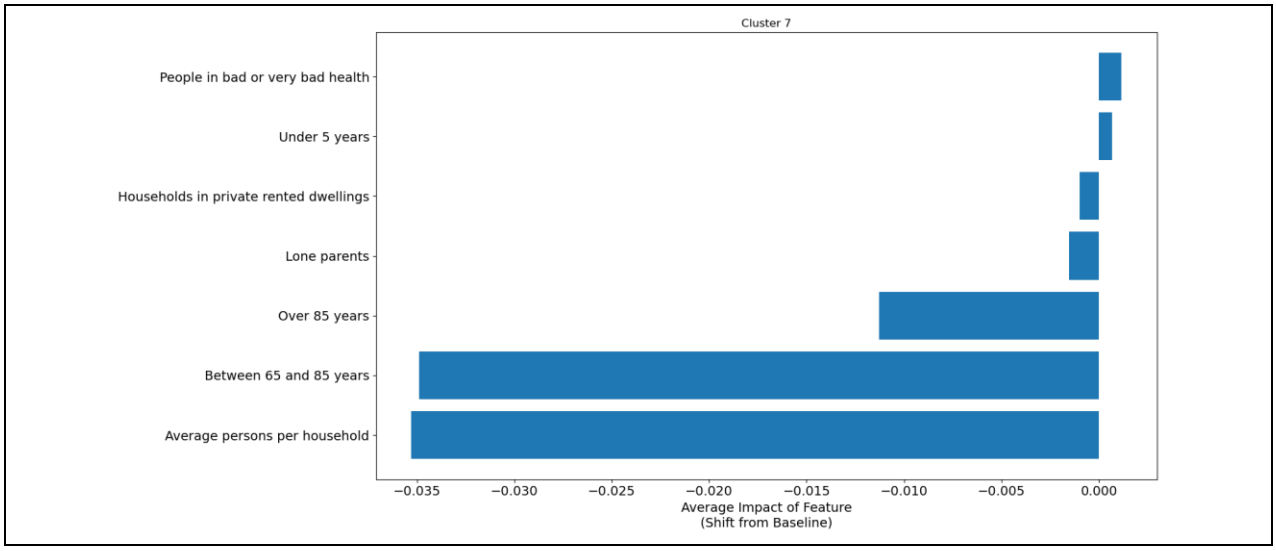


Figure 27: Visualisation of the 7 most influential features in each cluster. These are the 7 features that have the highest average absolute SHAP value. Here vulnerability is defined using the estimated PSR eligibility as computed by SSEN.

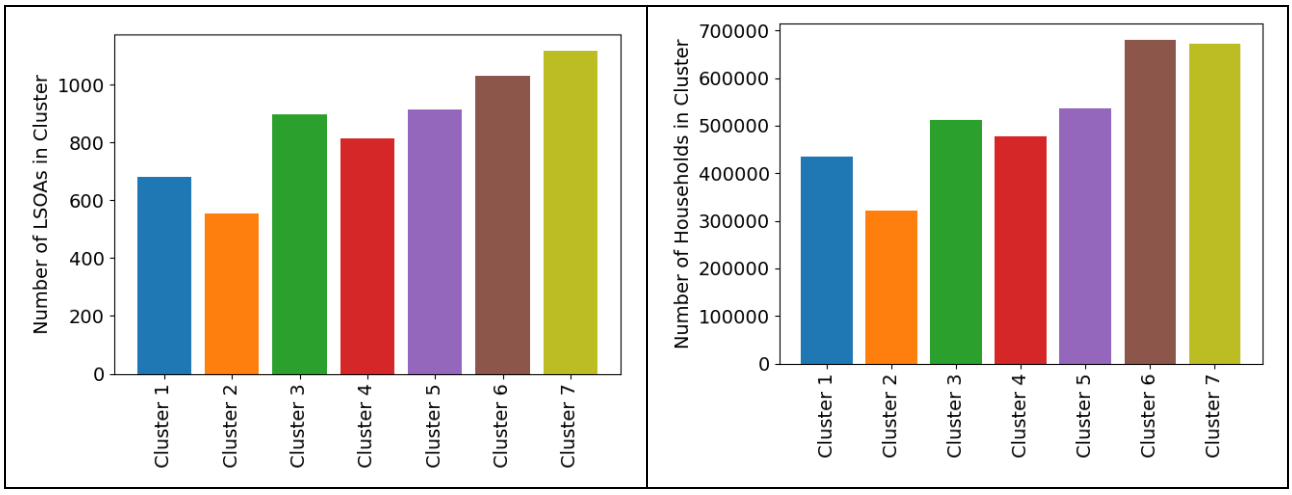


Figure 28: Cluster sizes, detailing how many LSOAs fall into the discovered clusters (left) and how many households lie in each cluster (right). Cluster 2 is the smallest, encompassing around 500 LSOAs, whereas cluster 7 is the largest, encompassing over 1000. Here vulnerability is defined using the estimated PSR eligibility as computed by SSEN.

.

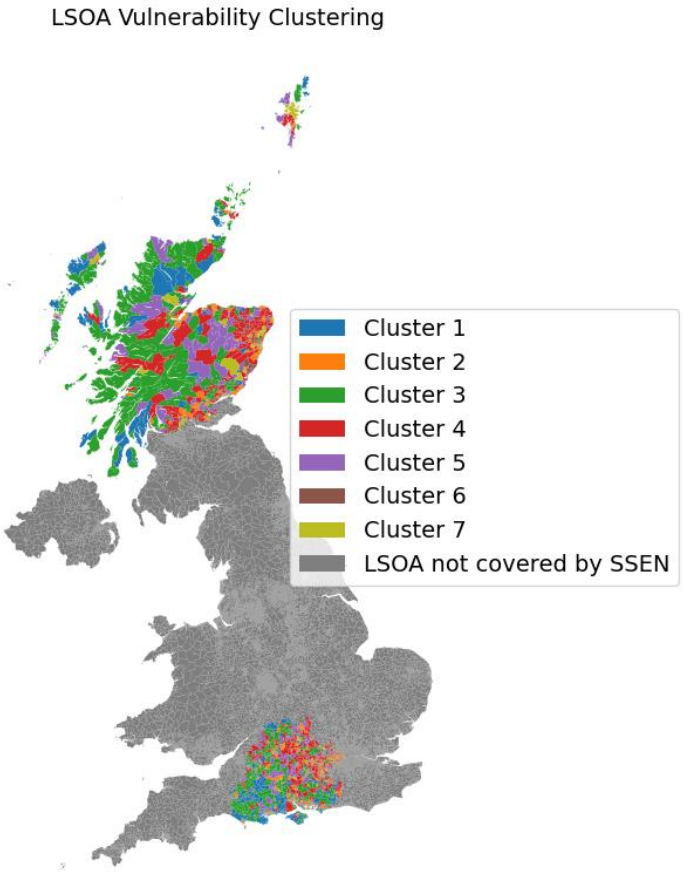


Figure 29: Map of the UK, with LSOAs coloured by the cluster they are a member of. LSOAs that are not serviced by SSEN are coloured in grey. Here vulnerability is defined using the estimated PSR eligibility as computed by SSEN.

Discussion of differences in results given different targets

Comparing the results in this section to that in the main body of the report, there are differences in the drivers of vulnerability one discovers when using the ***estimated*** eligibility of households for the PSR (computed by SSEN) compared to the ***actual*** number of PSR records. This is to be expected, as changing the definition of vulnerability the model learns will result in different features being picked as more or less important by the model, and hence the explanations computed from this model to determine drivers of vulnerability will be different. We discussed in section 1.1.1 the issue with using the estimated quantity in the computation of vulnerability: the machine learning model we apply may just re-learn the criteria SSEN use to generate the estimate, rather than the underlying drivers of vulnerability.

If we compare specific drivers of vulnerability identified in Figure 27, we see that ‘Under 5 years’ is present in multiple clusters, whereas this was not present at all in the results found in Figure 14 (where the actual PSR data is used to determine vulnerability). This highlights a fundamental difference between the two sets of results. The estimate produced by SSEN may consider customers eligible to be on the PSR if they have young children – indeed this is listed as one

eligibility criteria from Ofgem[3]. Therefore, having young children does indeed make a household eligible to be on the PSR, but when we look at the actual PSR records, this is not a significant driver of vulnerability. This does not mean any aspect of this process is incorrect: SSEN are likely correctly including this in an estimate of vulnerability, and this agrees with Ofgem, but equally, the actual PSR as it stands today underrepresents this group of vulnerable customers.

A further observation from Figure 27 is that almost all clusters are strongly driven by age, both young and old, and household size. It appears that age is a stronger driver of this estimate of vulnerability than it is in the actual PSR membership. This might be due to the fact that anyone who is of state pension age is eligible for PSR membership. However, many may either be unaware of the PSR, or feel they are personally 'not vulnerable' when they hear of its existence. Clearly, adjustments for eligibility will need to be applied, and the results in this section suggest that the adjustment needed for this reason is larger than those needed for other eligibility conditions. It may be that the SSEN methodology for estimating vulnerability levels corrects for age more so than it does for other categories, by necessity based on who is underrepresented on the PSR.

It is important to understand that keeping an up to date and accurate account of customers who are vulnerable is important to the wider operation of any DNO. However, we have seen here that there are significant differences if one performs analysis of the actual PSR records, compared to doing the same analysis on estimated vulnerability levels. Neither of these approaches is inherently wrong, however significant thought and care needs to be given when understanding how exactly one defines vulnerability, and how this impacts the results their analysis will yield.

[3] See https://www.ofgem.gov.uk/information-consumers/energy-advice-households/getting-extra-help-priority-services-register (Accessed on 20/10/2022).