Predicting sense of community and participation by applying machine learning to open government data

Alessandro Piscopo
University of Amsterdam[*]
Science Park 904, Amsterdam
The Netherlands
alessandro.piscopo@gmail.com

Ronald Siebes
VU University Amsterdam
De Boelelaan 1081a,
Amsterdam
The Netherlands
ronny@cs.vu.nl

Lynda Hardman
Centrum Wiskunde &
Informatica
Science Park 123, Amsterdam
The Netherlands
lynda.hardman@cwi.nl

ABSTRACT

Community capacity is used to monitor socio-economic development. It is composed of a number of dimensions, which can be measured to understand the possible issues in the implementation of a policy or the outcome of a project targeting a community. Measuring community capacity dimensions is usually expensive and time consuming, requiring locally organised surveys. Therefore, we investigate a technique to estimate them by applying the Random Forests algorithm on secondary open government data. Our research focuses on the prediction of measures for two dimensions: sense of community and participation. The most important variables for this prediction were determined. The variables included in the datasets used to train the predictive models complied with two criteria: nationwide availability; sufficiently fine-grained geographic breakdown, i.e. neighbourhood level. The models explained 77% of the sense of community measures and 63% of participation. Due to the low geographic detail of the outcome measures available, further research is required to apply the predictive models to a neighbourhood level. The variables that were found to be more determinant for prediction were only partially in agreement with the factors that, according to the social science literature consulted, are the most influential for sense of community and participation. This finding should be further investigated from a social science perspective, in order to be understood in depth.

Categories and Subject Descriptors

H.2.8 [**Database management**]: Database Applications—*Data mining*; I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*Learning*; J.4 [**Social and behavioral sciences**]: Sociology

General Terms

Algorithms, Measurement, Social Sciences

Keywords

Social dimensions, open government data, data science

1. INTRODUCTION

Community-based approaches are widely employed in publicly or privately funded programmes targeted to the promotion of socio-economic development and to address issues affecting disadvantaged neighbourhoods. Several of these approaches focus on building the capacity of a community, or community capacity (CC), either as a means to reach a certain goal, or as a goal in itself. CC is the ability of people in a community to act individually or collectively to undertake an action that will benefit the community itself [11]. It is used mainly in the implementation of public health policies, with applications also used in several other fields [16], such as tourism.

Whereas many definitions of CC can be found in the literature, they all agree that it is composed of several dimensions [21]. Those included in the majority of definitions are [11]: learning opportunities and skills development; resource mobilisation; partnership/linkages/networking; leadership; participatory decision making (or participation); asset-based approach; sense of community; communication; development pathway. Any change in these dimensions affects the capacity of a community [21]. Therefore, for any intervention targeting CC, it is important to measure its effects, to understand which dimensions are deficient and which initiatives should be taken to improve them [21].

Since high levels of CC increase the possibility of policies targeting a community to be successful [9, 21], the evaluation of CC dimensions facilitates policy makers and local administrators in understanding which issues might affect any planned initiative, the possible strategies to address them and the possibilities of success. Nevertheless, if measures of CC are not already available, obtaining them is generally too onerous for local institutions. The method usually followed to gauge CC is to organise local surveys [13], which may not be feasible due to their high costs. This also hampers the realisation of a longitudinal measurement of CC, which results in "lack of guidance on the relative importance of domains [or *dimensions*], the feasibility and benefits of long-term assessment of capacity building, the relationship between domains over time and to what extent measures of capacity

[*]MSc. Information Studies, Human Centered Multimedia. Research carried out in collaboration with the Centrum Wiskunde & Informatica, Amsterdam.

This research was carried out as a master thesis project within the MSc Information Studies, Human Centered Multimedia of the University of Amsterdam (http://gss.uva.nl/masters-programmes/content31/human-centered-multimedia.html).

development can be associated with health outcomes" [11, p. 3]. The absence of such measurements is reflected in a greater focus of the literature on the description of the process of CC building, rather than on its measurement [11].
A less resource-demanding method to measure CC dimensions would enable administrators to gain quickly and inexpensively an understanding of the characteristics of local communities, in cases in which organising a local survey is not feasible. In addition, it would raise the self-assessment ability of communities themselves and improve the accountability of local administrations. Moreover, it would be an instrument for researchers to perform a longitudinal study of CC dimensions on a larger scale.
An alternative method to obtain measures of CC dimensions relies on the results of national surveys on social aspects of the communities. Nevertheless, these surveys are often based on samples that are reliable at a national level, but do not involve a sufficient number of participants at a local scale. We do not investigate this method.
Another approach, investigated in this research, applies predictive algorithms to secondary data. With secondary data, we refer to data collected primarily for other purposes, which refers topics other than social dimensions, such as demographics or socio-economic data. This strategy does not require a large number of resources to supply measures of CC dimensions, as it takes advantage of data already available. Furthermore, in England – the context of our research – these data are available for the general population, which avoids uncertainties due to sample size. Our research investigated this approach with regard to two CC dimensions: sense of community and participation.

Research question

The main question that our research poses is: to what extent can we predict measures of participation and sense of community through applying a machine learning algorithm to secondary data? The measures obtained have to be theoretically suitable for use in the context chosen, and therefore have to comply with two criteria: consistent nationwide applicability, which means that the measures had to be available for any area within the context of our study; high geographic precision, i.e. they must be detailed at neighbourhood level [5]. As a secondary research question, we wanted to determine which variables had the highest influence for predicting sense of community and participation in the context chosen and whether they were in agreement with those determined using other models in the literature.

Structure of the paper

The paper is structured as follows. We first introduce the context of our study (Section 2). Subsequently, we present the related work followed with regard of the selection of social dimensions indicators and of the choice of the predictive algorithm (Section 3). Section 4 describes the method used for selecting the relevant variables for the models, as well as the data gathering and processing. This includes the tuning of the machine learning algorithm, the criteria for assessing its performance and determining which variables contributed the most to the predictions made. The data collected are described in Section 5 and results are presented in Section 6. Finally, strengths and limitations of our study are discussed in (Section 7) and conclusions are in (Section 8).

2. PROJECT STENTOR

The UK has released a wealth of open government data, made available in machine-readable formats, published according to open standards and released under an open license [19]. Notwithstanding these efforts, there are still issues concerning the full accessibility of datasets, which are often scattered over several departments. This might make arduous to retrieve the relevant datasets for a specific topic. Project Stentor was conceived to address issues connected with the accessibility of government data[1]. It is a UK project, carried out by two companies, MastodonC, specialised in big data analysis and applications, and Social Life, whose activities are related to community sustainability and development. The aim of Project Stentor is to create a platform to enable local administrators and policy makers to access, compare and analyse datasets from different sources, in order to gain new insights on a wide range of topics, from community dynamics to environmental issues.
Our research was performed in collaboration with the companies involved in Project Stentor. This collaboration covered the selection and the identification of the relevant sources and the exchange of information along all the research process. The platform developed within Project Stentor includes measures of social dimensions, created by Social Life on the basis of UK national surveys. These measures are matched to the Index of Multiple Deprivation decile or the Output Area Classification to which each area belongs. However, they do not provide values related to the single neighbourhoods, but only to determined typologies of areas. With our study, we aim at investigating the possibility to develop measures for sense of community and participation that are easy to obtain and matched to single neighbourhoods. We believe that such measure would be a valuable contribution to the Project Stentor and other similar projects.

3. RELATED WORK

We explain the criteria used to select the relevant variables for sense of community and participation and the machine learning algorithm used.

3.1 Social dimensions indicators

The first step to build our predictive models was to select the variables to include in each of them. Since our aim was to predict CC dimensions using secondary data, we needed to identify the datasets relevant for each dimension studied. Predictive models of social dimensions are generally built on a selected number of relevant indicators, or predictors, that are mapped to their appropriate measures [12, 13, 18, 20]. Beyond predicting the level of a social dimension, e.g. sense of community, the models generally aim at describing the relationships among that and the indicators used [12, 18], or at building an index that provides a measurement of a concept, by using proxy (*secondary*) data [20]. The indicators selection can be performed by assessing their relevance on the basis of theoretical assumptions [6, 12], which are confirmed or contradicted by an analysis of the data collected – often in a survey organised specifically for the study. Another approach is to submit the indicators selected from a literature review to the judgement of a group of experts, who have to assess the suitability of the indicators chosen for the con-

[1] *Project Stentor: Giving city data a voice.* Project prospectus, 24 October 2013.

text of the study [13]. These approaches were not feasible for our research, since it was out of our scope of our models to explain which factors influenced the social dimensions chosen and how and we were not able to submit our indicators list to any team of experts, due to time constraints. These selections included direct measurements of social dimensions among their indicators, which were collected using surveys made on population samples. Conversely, we wanted to use only secondary data, such as demographics and socioeconomic data, collected on the overall population. However, we used these studies to perform a first selection of the indicators for participation and sense of community. This selection was restricted by compiling a “wishlist”, in which each concept was connected with the measures that possibly described it, and identifying appropriate data sources and datasets, according to the method followed in [20] for the creation of an index for community resilience using secondary data. Differently from this method, our procedure did not include a further reduction of the indicators on the basis of the degree of correlation among them.

3.2 Prediction techniques

The studies mentioned in 3.1 use standard statistics to build their models[2], which contrasts with data mining, in the different focus on prediction accuracy. Table 1 provides an overview of the main differences among these two approaches.

Because of the strong assumptions formulated on the structure underlying the data [4], standard statistical techniques are more suitable to illustrate the relationships among the input variables and their relative importance. However, since they have to rely on domain knowledge – i.e. a theoretical framework set by experts in the field – they face the risk of drawing conclusions concerning more the theory adopted, rather than the data itself. Furthermore, domain experts – social scientists, statisticians – are needed to build a model. On the other hand, data mining requires only limited domain knowledge and predicts outcome variables by discovering patterns inherent to the data [7]. The output of data mining techniques is therefore less subject to the risk of relying on an erroneous theory. On a more practical side, they can be applied more easily by experts of other disciplines and deployed on a larger scale, due the reduced role of domain expertise [3]. This is in accordance with our purpose of building a predictive model suitable to be used by several types of figures interested in measuring CC dimensions.

	Standard statistics	Data mining
Example techniques	Linear regression, factor analysis, ANOVA.	Neural networks, decision trees, SVM.
Domain knowledge	Based on strong theoretical assumptions.	Relying on limited domain knowledge.
Information on data structure	Detailed information on the relationships among variables involved.	Little information on the relationships among variables.
Model validation[4]	Goodness-of-fit tests, residual examination.	Prediction accuracy.

Table 1: Main differences between standard statistics and data mining [4, 7].

One of the issues of data mining techniques is that they are often considered as “black boxes”, in that they provide little interpretable information about how variables determine the final prediction. For example, the predictions made by Support Vector Machines (SVMs), one of the most accurate learning models [24], are difficult to explain [2]. Not all of these techniques have such interpretability problems. Random Forests offer clear insights about the predictive importance of the variables included in the model [23], while providing high prediction accuracy, compared with other algorithms [24]. This technique, applied already to several fields, such as genetic, bio-informatics and, in the social science field, psychology and organisations management and sustainability [10], is suitable for both classification and regression tasks[3]. The characteristics of the Random Forests algorithm, which grows successive decision trees, using a random sample of the training data for each of them, make it robust to overfitting [22] and avoid the problems derived by the “multiplicity of good models”. This definition refers to the possibility of building a high number of equally predictive models in the presence of highly dimensional datasets, by removing even small subsets (2 to 3%) [4]. Moreover, Random Forests is suitable for training data with a small number of instances (n) and a large number of variables (p), even in extreme cases in which $n \ll p$ [23].

Another advantage of Random Forests concerned the quality of the variable importance measure provided. The most reliable of the built-in variable importance functions in this algorithm is the “permutation accuracy importance” [23]. It computes the importance value of a variable by randomly permuting it, calculating the prediction accuracy before and after each permutation and averaging this difference over all the trees. This importance measure has been shown to be both stable – among different iterations of the algorithm – and able to convey “the importance of variables in interactions too complex to be captured by parametric regression models” [23, p. 324].

The high prediction accuracy and the interpretability of the results of the Random Forests algorithm were suitable for the creation of a predictive model to be used in real settings and the investigation of the most relevant variables for prediction. Furthermore, other characteristics – i.e. robustness to overfitting and to the multiplicity of good models problems, suitability for datasets with many variables and few instances – were appropriate for the datasets created, as these had a large p (about 50 variables) and small n (about 300 instances). Therefore, we chose to use Random Forests.

4. METHOD

We explain the criteria for the choice of the CC dimensions studied then we illustrate how we selected, collected and processed the data. The choice of CC dimensions and the data selection proceeded in parallel, so their outcomes influenced one another.

4.1 Selection of community capacity dimensions

[2] Unless differently specified, the terminology adopted in this subsection follows closely [7].

[3] With classification task, we refer to a function whose purpose is to predict a categorical outcome value, whereas the goal of a regression task is to predict a continuous value.

We chose to investigate only two CC dimensions: sense of community and participation, based on the availability of measures to be used as dependent variables for training our predictive models. The available measures had to satisfy three requirements.

- They had to match the social dimensions investigated in our study as closely as possible. From a first overview of the available data, none had been collected with the explicit purpose of measuring CC dimensions, so the matching might not be exact.
- They needed to have a consistent national coverage. In the UK the same statistics geography is used for England and Wales, whereas there are some differences in the ones used for Scotland and Northern Ireland. Therefore, the maximum coverage possible was England and Wales.
- Their geographic detail had to be able to provide information about a small to medium-sized neighbourhood (up to a few thousand residents). Using the nomenclature of the UK Office of National Statistics (ONS) geography, which was employed with minor changes in the 2001 and 2011 censuses, the best geographic breakdown for this purpose was the Lower Super Output Area (LSOA). Its level of detail is appropriate to describe a neighbourhood, while it provides wider availability than the immediately smaller ONS statistical subdivision, the Output Area (OA, see Table 2 and Figure 1 for further details). Another advantage of using data related to smaller areas is that each measurement represents an instance of the dataset used to train our model, therefore smaller areas provide a higher number of instances.

Notwithstanding the wide availability of national surveys investigating social dimensions in the UK, such as one of the most comprehensive, the United Kingdom Household Longitudinal Study (UKHLS) or Understanding Society survey, we were unable to use them, because of the long times to access the data.

Therefore, we used the National Indicators NI 002[4], to measure sense of community, and NI 003[5], to measure participation, whose geographic breakdown is the local authority (LA) level. A definition of these social dimensions and more details about the related measures are given in the next subsection.

Geography	Avg. no. residents	Avg. no. households	Total no. of areas*	Avg. units per higher level
OA	309	129	181,408	5-7
LSOA	1,614	672	34,753	7-9
MSOA	7,787	3,245	7,201	–

Source: ons.gov.uk

*In England and Wales, 2011.

Table 2: Office of National Statistics Geography details. Considering the extension of the areas and the availability of the data, LSOA was the most suitable level to provide measures related to neighbourhoods.

4.1.1 *Sense of community and participation measures*

Sense of community includes several elements, joined together in the following definition: "sense of community is a feeling that members have of belonging, a feeling that members matter to one another and to the group, and a shared faith that members' needs will be met through their commitment to be together" [14, p. 9]. It plays an essential role in CC building, as it increases the active membership at the basis of participation, influences the collective norms and values and improves the mobilisation of resources [9]. As already mentioned, the measure used to train our model for sense of community was NI 002 (*% of people who feel that they belong to their neighbourhood*), which is constructed on the basis of the responses to the question "How strongly do you feel you belong to your immediate neighbourhood?", by calculating the ratio among the number of positive answers ("fairly strongly" or "very strongly") and the total of valid ones. Although it does not describe all the aspects of sense of community, we used NI 002, since it was the closest measure available.

Participation is defined as the "people's engagement in activities within the community" [16]. It is an essential quality of CC, as community members may gain an understanding and act on issues concerning the community as a whole only by participating in small groups or smaller organisations [9]. Participation is strongly linked to other CC dimensions as it is needed by local leaders in managing activities for the community and provides a base for skills and resources [9]. The measure used to provide values for participation was NI 003 (*Civic participation in the local area*) was used to provide values for participation, as it provided an appropriate measure for this social dimension. NI 003 is built using the positive answers to a question about whether the respondents had taken part in any group – from a list of different types of groups – making decisions affecting their local area and not related to their profession, in the previous 12 months.

The geographic breakdown of NI 002 and NI 003 is the local authority (LA) level, their coverage is the whole of England. Since they provide a measure for each LA in this country, the total number of values for each of them is 353 (for 354 LAs, one value is missing). The responses on which they are built were collected within the 2008 Place Survey, which is now discontinued. This survey was administered by local authorities and "provides information on people's perceptions of their local area and the local services they receive"[6]. It used a multi-stage stratified random sample of a minimum size of 1,100 addresses of adults resident per LA, for a total of 518,772 individual participants nationwide. Both the measures provide continuous values, with higher ones indicating better performance, i.e. higher levels of sense of community or participation.

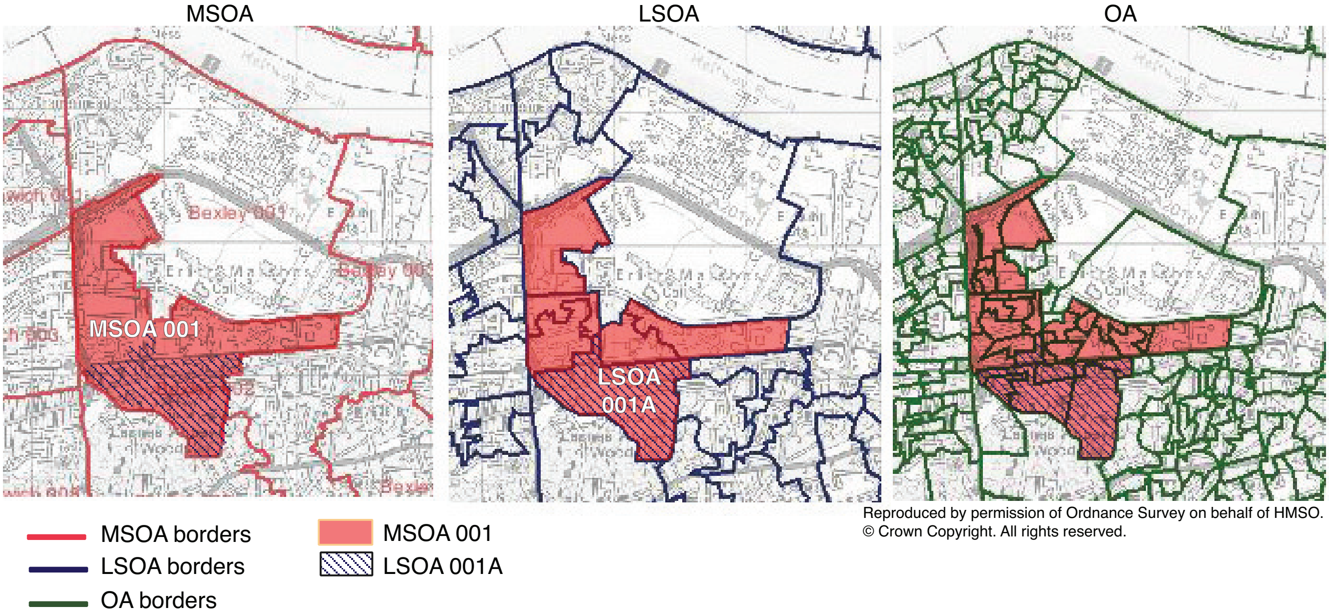
4.2 Data gathering and processing

4.2.1 *Data selection criteria*

[4] http://data.gov.uk/dataset/ni-002-percentage-of-people-who-feel-that-they-belong-to-their-neighbourhood.

[5] http://data.gov.uk/dataset/ni-003-civic-participation-in-the-local-area.

[6] http://discover.ukdataservice.ac.uk/catalogue/?sn=6519.



Reproduced by permission of Ordnance Survey on behalf of HMSO.
© Crown Copyright. All rights reserved.

Figure 1: Sizes of Output Areas (OA) and Super Output Areas (LSOA, MSOA). Larger areas are aggregations of smaller ones.

We selected variables for our models on the basis of the relevant indicators of participation and sense of community. For each indicator, a hypothesis about the measures describing it more appropriately was made [20], compiling a wishlist of variables to be included. Subsequently, we checked which variables in the datasets available from the open government data sources selected[7] matched the ones in the wishlists and we adapted these. For example, social networks may be relevant indicators of participation [6]. Therefore, we first built a wishlist of measures for social networks based on the variables used in [6]. These included the presence of family and friends in the neighbourhood and the frequency of relations with the neighbours. Nonetheless, these measures were either not available from the data sources selected, or the variables available did not meet the requirements set. However, we considered that the number of people providing unpaid care and the percentage of people working in the neighbourhood[8] might provide an indirect measurement of social networks, therefore we included them in our models. In order to be suitable for selection, datasets had to comply with three criteria: geographical coverage, geographical detail and time (see Table 3 for an overview of the data selection criteria). Geographical coverage and detail were related to the requirements stated for the measures we wanted to obtain and to the characteristics of the dependent variables available: data had to be at nationwide coverage, i.e. England, since this was the coverage of the measures used for participation and sense of community; they had to be applicable to small neighbourhoods. We kept this latter condition in order for our models to be theoretically suitable for smaller areas, although they were trained on local authority level data, The other criterion, time, required that data were available for a time span as close as possible to the dependent variables. Finally, we discarded the indicators for which no measures were available.

4.2.2 *Data cleaning and preparation*

The datasets collected contained no missing values or rogue

Criterion ***Condition sought***	**Condition available** **• Notes**
Geographical coverage *Nationwide coverage*	England • Whereas the datasets selected were all available for England and Wales – which would have provided a bigger training set –, the measures of the social dimensions were available only for England.
Geographical detail *Neighbourhood level*	LA • LSOA level was the one that provided the best combination of geographical detail and availability. However, we used LA level data, as this was the most accurate level of detail for the measures of sense of community and participation.
Time *Closeness to social dimensions measures used*	2008-2011 • The measures of sense of community and participation were referred to 2008. Given the long evolution times of social dimensions [20], we decided to include data up to 2011, the year in which the last UK census on the general population took place.

Table 3: Data selection criteria. The characteristics of the social measures available determined the characteristics of the data used for prediction, making them differ from the optimal ones.

[7] See Appendix A for a list of the sources used.

[8] For more details about these variables, see Appendices B and C.

attributes, since they complied with the quality standards of the Office of National Statistics and other government departments, i.e. accuracy, coherence and comparability. The variables depending on the local authority size were normalised, dividing them by the total number of units to which they referred, e.g. number of residents or number of households. Data related to the ethnic composition of the population were used to calculate ethnic fragmentation (see Appendices B and C for more details), which is correlated with participation and social cohesion [1].

4.2.3 Data processing

The aim of our study was to build models to predict levels of sense of community and participation. After selecting the variables to be included in the two models, we applied to them the machine learning technique chosen. Both for sense of community and participation the values to be predicted – the dependent variables – were continuous, therefore the Random Forests algorithm had to be applied to a regression problem. This algorithm provides a measure of its prediction accuracy based on a random sample of the training data, called out-of-bag (OOB) sample, left out for each tree grown. This sample is used for the evaluation of the single trees, and the accuracy of the whole model is calculated by averaging the results of all the trees. Because of these characteristics, separate training and test sets were not needed. We applied Random Forests using one of its R implementation, the package *party*[9]. This package was chosen because of its reliability the importance of variables, even when these are highly correlated [24].

In order to optimise the prediction accuracy and the stability of the model, we tuned the algorithm used, by setting two parameters, *mtry*, i.e. the number of variables randomly chosen at each split, and *ntree*, i.e. the number of trees in the forest [8]. Finding the optimal settings for these parameters is also important to lower the bias in the selection of important variables [24]. For each model, we tuned the algorithm by setting *ntree* and *mtry* to their default values for regression ($ntree$ = 500, $mtry = p/3$), increasing them by 100 (*ntree*) and by 5 (*mtry*), until we could not observe any improvement in the prediction accuracy. This was assessed by the mean squared error (MSE) and the R^2, calculated on the OOB sample (i.e. for MSE, lower is better; for R^2, higher is better). R^2, called coefficient of determination, is a measure of how a regression model fits the variability of a data set. It is described by the formula $R^2 = 1\text{-}\frac{SS_E}{SS_T}$, where SS_E is the sum of squared errors and SS_T is the total sum of squares. The accuracy measures of the models (MSE and R^2) trained with the optimal *mtry* and *ntree* were evaluated by comparing them to predictive models of social dimensions found in the literature, to better assess their performance for a possible use in a real setting.

The variable importance was computed[10] accounting for the conditional importance of the variables. We assessed the results relative to the predictivity of the variables by observing how each variable ranked among the others. We did not report the importance values produced by the algorithm, since these are not comparable among different studies [23]. However, in order to better convey the degree of predictivity of each variable with respect to the others, we provided the ratios among their importance values.

Finally, in order to reduce the number of variables included in the model, we wanted to identify which ones were irrelevant for prediction. To do that, we followed a heuristic, where variables can be considered informative and important if their importance score is above the absolute value of the variable with the lowest negative score [23]. This heuristic relies on the fact that irrelevant and uninformative variables present importance values randomly varying around zero.

5. DATA DESCRIPTION

A total of 23 datasets were collected[11], the majority of them (17) from the 2011 Census. These include Key Statistics (KS) and Quick Statistics (QS), which both cover the full range of census topics, with the difference that the former ones provide summary figures, such as ratios over the overall sample and combinations of several variables, whereas the latter ones include the most detailed information on a single topic[12]. QS provide the maximum possible detail (OA), whereas KS are often available only for LSOAs and MSOAs. The indicators selected covered various areas, such as socio-economic characteristics, socio-demographics and housing conditions.

The datasets related to sense of community and participation are shown in tables 4 and 5. Both the datasets created for our predictive models had 316 instances, each instance representing an English local authority. The difference among the number of values of NI 002 and NI 003 and the final number of instances in the datasets was due to divergences between the administrative geographies used in some datasets. Therefore, not all of the English local authorities were included in the datasets. The variables included in each dataset are listed in detail in Appendix A and B. However, this is a summary of their characteristics:

- The sense of community dataset had 48 continuous independent variables and one continuous dependent variable (NI 002) (See Fig. 4 and Appendix B). This had a maximum value of 75.1 and a minimum one of 42.8.
- The participation dataset had 48 continuous independent variables and one continuous dependent variable (NI 003) (See Fig. 5 and Appendix C; the equivalence of the number of variables in the two datasets is accidental). This had a maximum value of 25.7 and a minimum of 7.6.

6. RESULTS

Sense of community

The optimal settings for the sense of community model were *mtry* 44 and *ntree* 1,000. Using these values, the model yielded an MSE of 9.5 and an R^2 of 76.5% (Fig. 2). The prediction accuracy did not increase by growing further trees or raising the number of variables chosen at each split, if not

[9] R version 3.1.0, on Mac OS 10.7.5; *party* package version 1.0-15.

[10] We used the command *varimp* from the *party* package.

[11] See appendix for a detailed list of the variables included from each dataset.

[12] http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/table-types/index.html.

Category	**Indicators**	**No. of datasets (year)**	**Source datasets**	**No. of variables used***
Socio-demographics	Gender, median age, length of residence in the UK, ethnic fragmentation, religion.	5 (2011)	2011 Census.	34 (16)
Socio-economic characteristics	Employment sector, income, level of qualification.	2 (2011), 1 (2010)	2011 Census, English indices of deprivation 2010, Benefits claimants.	7 (7)
Health	Health conditions.	1 (2011)	2011 Census.	2(2)
Households composition	Number of households with children, married couples, civil partnerships, not living in a couple.	2 (2011)	2011 Census.	7 (7)
Tenure and housing category	Homeowners, tenants.	1 (2011)	2011 Census.	3(3)
Social networks	People providing unpaid care in the neighbourhood, people working in the neighbourhood.	2 (2011)	2011 Census, Core accessibility indicators.	4 (4)
Resources and environment	Religious organisations, education facilities, pollution, town centres accessibility, commercial centres accessibility, property crimes, crimes against the person.	5 (2011), 1 (2010)	2011 Census, Core accessibility indicators, English indices of deprivation 2010, data.police.uk.	7 (9)
Total		20		66 (48)

*In brackets, the number of variables included in the model after aggregation.

Table 4: Datasets collected for sense of community.

Category	**Indicators**	**No. of datasets (year)**	**Source datasets**	**No. of variables used***
Socio-demographics	Gender, median age, length of residence in the UK, ethnic fragmentation, proficiency in English.	6 (2011)	2011 Census.	28 (10)
Socio-economic characteristics	Employment status, women in employment, hours worked, income, people receiving benefits, socio-economic status, level of qualification.	7 (2011), 1 (2010)	2011 Census, English indices of deprivation 2010, Benefits claimants.	19 (19)
Health	Health conditions.	1 (2011)	2011 Census.	2 (2)
Households composition	Number of households with children, married couples, civil partnerships, not living in a couple.	2 (2011)	2011 Census.	7 (7)
Tenure and housing category	Homeowners, tenants, social housing share.	1 (2011)	2011 Census.	3 (3)
Social networks	People providing unpaid care in the neighbourhood, people working in the neighbourhood.	2 (2011)	2011 Census, Core accessibility indicators.	4 (4)
Resources and environment	Religious organisations, professional organisations, education facilities.	1 (2011)	2011 Census.	3 (3)
Total		21		66 (48)

*In brackets, the number of variables included in the model after aggregation.

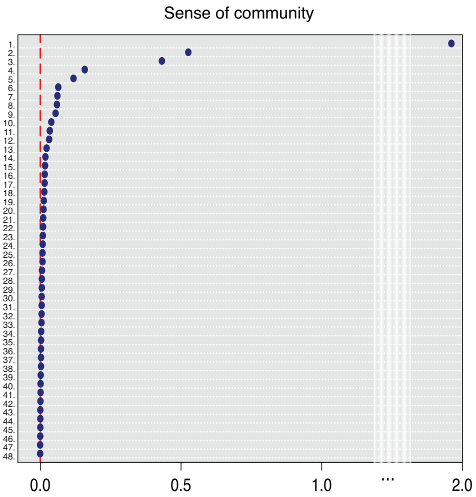
Table 5: Datasets collected for participation.

decreased slightly. According to the heuristic enunciated in 4.2.3, only 7 variables out of 48 could be regarded as not important for prediction (Figure 4). The median age of the population was the most predictive variable, followed by the share of people providing 1 to 19 hours unpaid care a week (importance value ratio compared to the higher ranking variable: 0.27) and by the index of work accessibility (0.82). The share of people in intermediate occupations (0.36) and the number of violent crimes (0.75) ranked in the fourth and fifth positions.

Participation

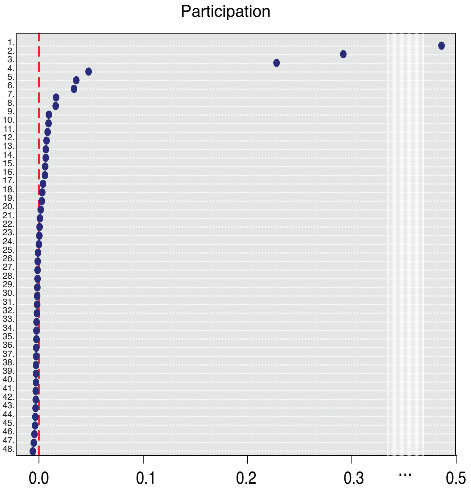
The optimal settings for the participation model were *mtry* 27 and *ntree* 1,100, which yielded MSE 3.7 and R^2 62.5% (Figure 3). Growing further trees or increasing the number of variables at each split did not improve the accuracy of the model. According to the heuristic in 4.2.3, only 10 variables out of 48 could be defined as neither informative, nor important (Figure 5). The variable with the highest importance value was the proportion of people in intermediate occupations, followed by the proportion of people with a level 4 of education or higher (importance value ratio compared to the higher ranking variable: 0.51). The third variable was the share of small employers and own account workers (0.82), while the fourth and fifth ones were the percentages of households with cohabiting couples and dependent children (0.23) and of people of the same sex living in a couple, cohabiting or in a registered partnership (0.59).

7. DISCUSSION



1. Median age (KS102EW0019)
2. People providing unpaid care 1 to 19 hrs./week (QS301EW0003)
3. Work accessibility
4. People in intermediate occupations (KS611EW0006)
5. Violent crimes (rate over population)
6. Living in a couple, cohabiting (opposite sex) (QS108EW0004)
7. Food stores accessibility
8. Health condition: good health (QS302EW0003)
9. Residents in the UK, 5 to 10 yrs. (QS803EW0005)
10. Vehicle crimes (rate over population)
11. Religion: Jewish (QS208EW0005)
12. Residents in the UK, 2 to 5 yrs. (QS803EW0004)
13. Tenure: Owned (QS403EW0002)
14. Ethnic fragmentation
15. Burglary (rate over population)
16. Living in a couple, married (QS108EW0003)
17. One family household, cohabiting couple w/ dependent children (KS105EW0009)
18. Highest level of qualification: level 4 or above (QS501EW0007)
19. Religion: Buddhist (QS208EW0003)
20. Religion: Hindu (QS208EW0004)
21. Other household types, w/ dependent children (KS105EW0013)
22. Highest level of qualification: other qualifications (QS501EW0008)
23. Living environment score (2010 id. of deprivation)
24. Religion: other religion (QS208EW0008)
25. Tenure: Private rented (QS403EW0009)
26. Communal etablishments: other, education (QS420EW0027)
27. Share of females over the population
28. Schools accessibility
29. Religion: Sikh (QS208EW0007)
30. Residents in the UK, ≤ 2 years (QS803EW0003)
31. Highest level of qualification: level 3 (QS501EW0006)
32. Religion: Muslim (QS208EW0006)
33. Criminal damage
34. Religion: No religion (QS208EW0009)
35. Health condition: very good health (QS302EW0002)
36. Resident in the UK: ≥ 10 yrs. (QS803EW0006)
37. People providing unpaid care 20 to 49 hrs./week (QS301EW0004)
38. Religion: Christian (QS208EW0002)
39. One family household, married or same-sex civil partnership couple, w/ dependent children (KS105EW0006)
40. Not living in a couple (QS108EW0006)
41. Born in the UK (QS803EW0002)
42. Income score (2010 index of deprivation)
43. People providing unpaid care ≥ 50 hrs./week (QS301EW0005)
44. Tenure: social rented (QS403EW0006)
45. Never worked and long-term unemployed (KS611EW0011)
46. Communal etablishments: other, religious (QS420EW0032)
47. Town centres accessibility
48. Living in a couple, in a registered partnership or cohabiting (same sex) (QS108EW0005)

Figure 4: Variable importance for sense of community (the dashed line indicates zero; the names in brackets indicate the source dataset or, when this belongs to the 2011 Census, the original name of the variable). On the right, the names of the variables, ranked by their importance value; the values of the lower 41, out of 48, variables varied around zero.



1. People in intermediate occupations (KS611EW0006)
2. Highest level of qualification: level 4 or above (QS501EW0007)
3. Small employers and own account workers (KS611EW0007)
4. One family household, cohabiting couple w/ dependent children (KS105EW0009)
5. Living in a couple, in a registered partnership or cohabiting (same sex) (QS108EW0005)
6. Hours worked: ≥49 hrs./week (QS604EW0007)
7. One family household, married or same-sex civil partnership couple, w/ dependent children (KS105EW0006)
8. Tenure: Private rented (QS403EW0009)
9. Hours worked: 31 to 48 hrs./week (QS604EW0006)
10. Proficiency in English: Cannot speak English (QS205EW0006)
11. Economically active females: full-time employee (KS603EW003)
12. Work accessibility
13. People in lower managerial, administrative and professional occupations (KS611EW0005)
14. People in routine occupations (KS611EW0010)
15. People providing unpaid care ≥50/week (QS301EW0005)
16. Proficiency in English: Cannot speak English well (QS205EW0006)
17. People providing unpaid care 20 to 49 hrs./week (QS301EW0004)
18. Tenure: Social rented (QS403EW0006)
19. Resident in the UK: ≥10 yrs. (QS803EW0006)
20. Tenure: Owned (QS403EW0002)
21. Economically active: unemployed (QS601EW0009)
22. Other household types, w/ dependent children (KS105EW0013)
23. Highest level of qualification: level 3 (QS501EW0006)
24. People in lower supervisory and technical occupations (KS611EW0008)
25. People in higher managerial, administrative and professional occupations (KS611EW0002)
26. People in semi-routine occupations (KS611EW0009)
27. Economically active females: part-time employee (KS603EW003)
28. Communal etablishments: medical and care establishment (QS420EW0012)
29. Born in the UK (QS803EW0002)
30. Ethnic fragmentation
31. Highest level of qualification: other qualifications (QS501EW0008)
32. Not living in a couple (QS108EW0006)
33. People providing unpaid care 1 to 19 hrs./week (QS301EW0003)
34. Living in a couple, married (QS108EW0003)
35. Resident in the UK: 2 to 5 yrs. (QS803EW0004)
36. Full-time students (KS611EW0015)
37. Residents in the UK, ≤ 2 yrs. (QS803EW0003)
38. Median age (KS102EW0019)
39. Income score (2010 id. of deprivation)
40. Never worked and long-term unemployed (KS611EW0011)
41. Income claimants
42. Living in a couple, cohabiting (opp. sex) (QS108EW0004)
43. Communal etablishments: other, education (QS420EW0027)
44. Residents in the UK, 5 to 10 yrs. (QS803EW0005)
45. Health condition: very good health (QS302EW0002)
46. Communal etablishments: other, religious (QS420EW0032)
47. Share of females over the population (QS104EW0003)
48. Health condition: good health (QS302EW0003)

Figure 5: Variable importance for participation (the dashed line indicates zero; the names in brackets indicate the source dataset or, when this belongs to the 2011 Census, the original name of the variable). On the right, the names of the variables, ranked by their importance value; the values of the lower 38, out of 48, variables varied around zero.

Accuracy of the model and applicability

The sense of community model obtained the best results for explaining the variation of the dependent variable (see Fig.s 2 and 3). The higher MSE for this model can be related with the higher range of the sense of community measure. Neither of the models built was suitable to predict CC dimensions at neighbourhood level, as this required an LSOA geographic breakdown. Nevertheless, the results achieved are promising for future applications in real contexts, as they show that secondary data can be used effectively to predict the social dimensions studied, by applying machine learning on them. The prediction accuracy was high, compared to previous studies in which parametric models were used. As an example, the model developed by [15], which attempts to predict participation in community organisations in New York, Baltimore and Salt Lake City, explains 28% of the variance of participation at individual level and 52% at block level. The model built by [12] to predict sense of com-

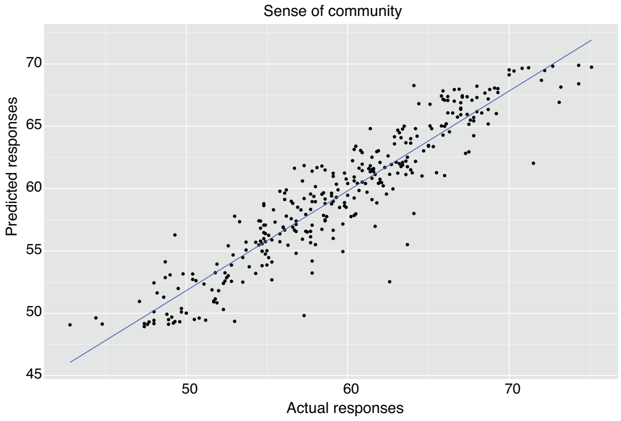


Figure 2: Sense of community (NI 002): plot of the predicted responses to the actual ones. The closer the predicted responses are to the line, the better the model fits the actual data (different scale from participation, Figure 3).

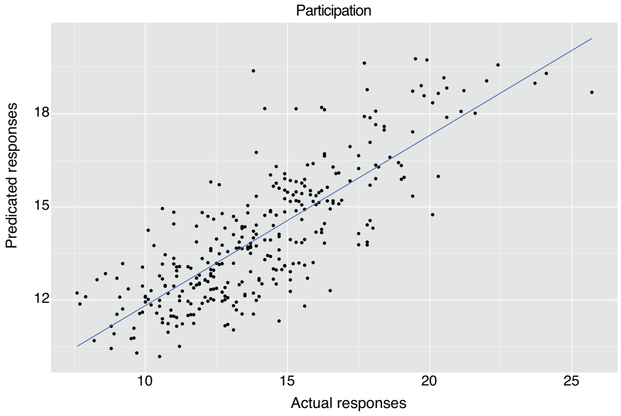


Figure 3: Participation (NI 003): plot of the predicted responses to the actual ones (different scale from sense of community, Figure 2).

munity in New York explained 39% of the variance of the outcome variable at individual level and 68% at block level. However, both these models include data from surveys organised on samples at local level, thereby not satisfying the requirements set for our study. Moreover, even though our models accounted for a higher percentage of the variance of the dependent variables in both cases, in order to provide a more valid comparison, a test of their accuracy on smaller areas is required. In order to do this, the most appropriate geographic breakdown is LSOA, which we have seen to be the level providing the optimal combination of availability and detail. However, the UK national surveys currently organised do not provide reliable data at this level, therefore locally organised surveys providing detailed information on CC dimensions are needed, to be used as ground truth for further studies.

Predictive variables

One of the strengths of our approach is the inclusion of a large number of variables, whereas other models, such as those mentioned in 3.2, rely on a narrower selection. This characteristic allowed to take into account also factors which are generally considered to have only a secondary effect on sense of community and participation, but that still may be helpful to improve a prediction of their measures.

The variables with the highest importance values were only partially in agreement with indicators found in the literature to be influencing participation and sense of community the most. Although Dekker [6, p. 370] concludes that "socio-economic status by itself has no positive or negative effect on participation", the proportion of people in intermediate occupations and the proportion of small employers and own account workers ranked at the first and third position among the most predictive variables for that social dimension. Furthermore, age of the population and ethnic fragmentation, both strong indicators of participation levels [1, 17], were not determinant for building the outcome value in our model. On the other hand, the level of education and the share of households with couples and children ranked high in our model, which agrees with the consulted literature [6, 17]. The importance of the share of people living in private rented houses may be seen in agreement with what stated by [6], if we consider it as a 'negative' of the proportion of owner occupiers. As for sense of community, [18] identifies the level of deprivation and the proportion of married people in the neighbourhood as the most important predictors, followed by "gender, age, household income, ethnicity and cohabitation with a partner." Of these, age and cohabitation (variables: median age and living arrangement: cohabiting (opposite-sex)) figured among the most important predictors also in our model. The importance of the length of residence in the UK, the percentages of homeowners and of people providing unpaid care in the neighbourhood may be associated with the relevance of place attachment and social networks in determining sense of community, as [12] reports. The role of vehicle and violent crimes in predicting sense of community is stated by [20], who include property crime rate among the indicators used to measure community bonds. Although a connection between religious faith and sense of community is highlighted by [18], we found no explicit mention of Judaism, whose number of adherents figured among the best predictors. Ethnic fragmentation did not rank among the highest predictive variables for the sense of community model.

However, "predictors thought to be important in a conventional model, may prove to be worthless in output from an ensemble analysis" (i.e. the typology of algorithms to which Random Forests belongs) and vice versa [3, p. 31], therefore the differences among the indicators of participation and sense of community found in the literature using conventional statistics and the one identified with Random Forests should be addressed under a social science perspective, in order to understand their meaning. Since the importance values provided by the Random Forests algorithm do not provide any description of how a variable influences the predicted outcomes, such research should also focus on explaining the relationships among participation and sense of community and the important variables highlighted in this research.

Time

CC dimensions are often measured to assess how they change during the implementation of a programme, such as in [13]. Since we used data collected over a long time span (2008-

2011), the measures provided by our models are not suitable for such purpose. The majority of the datasets we used are from the 2011 Census. Censuses in the UK are organised every ten years, therefore other data sources need to be found, in order to produce updated measures between one census and another.

8. CONCLUSION

We used Random Forests to build two models for predicting measures of sense of community and participation in English communities. These models yielded nationwide measures of both at local authority level, with high accuracy, compared to other models built using conventional statistics. The unavailability of data at a more detailed level for the dimensions studied did not allow the constructions of models to predict neighbourhood level measures. Further work to build more geographically accurate models should then rely on other sources, such as locally organised surveys. In addition, one of the reasons for the lack of more geographically detailed data regarding sense of community and participation is the bureaucratic process to connected to data disclosure policies. Because of this, we believe that further efforts are required from government authorities to increase the accessibility of government data, by implementing faster procedures to request data covered by privacy related restrictions.

Other achievements of our study were the identification of datasets containing measures related to the indicators of sense of community and participation found in the literature and the selection of predictive variables for these two dimension using Random Forests. About the latter ones, further research should address the differences among these variables and the indicators suggested by previous studies to better understand them and explain the relationships among the most predictive variables and the dimension predicted. Finally, further study should evaluate a fully data-driven approach, which would make a selection of the variables in the predictive models regardless of any domain knowledge. All the variables complying with the geographic and temporal requirements enunciated in 4.2.1 should be included in the models. Successively, their number would be narrowed down by using a feature of the Random Forests algorithm, which allows to eliminate the variables that are irrelevant for prediction. Using this method, the selection would be made only on the basis of the importance values generated by the algorithm, i.e. of the predictivity of the variables.

Acknowledgements

We thank MastodonC and Social Life, specifically Francine Bennett and Saffron Woodcraft, for their support. Furthermore, thanks go to Arjen de Vries and the Information Access group at the Centrum Wiskunde & Informatica, for their advice and the many illuminating conversations.

9. REFERENCES

[1] Alesina, A., and La Ferrara, E. Participation in heterogeneous communities. *The Quarterly Journal of Economics 115*, 3 (2000).

[2] Barbella, D., Benzaid, S., Christensen, J., Jackson, B., Qin, X., and Musicant, D. Understanding Support Vector Machine classifications via a recommender system-like approach. In *DMIN* (2009), pp. 305–311.

[3] Berk, R. An introduction to ensemble methods for data analysis. *Sociological Methods & Research 34*, 3 (2006), 263–295.

[4] Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science 16*, 3 (2001), 199–231.

[5] Chainey, S. Identifying priority neighbourhoods using the vulnerable localities index. *Policing 2*, 2 (2008), 196–209.

[6] Dekker, K. Social capital, neighbourhood attachment and participation in distressed urban areas. a case study in The Hague and Utrecht, the Netherlands. *Housing Studies 22*, 3 (2007), 355–379.

[7] Friedman, J. Data Mining and Statistics: What's the connection? *Computing Science and Statistics 29*, 1 (1998), 3–9.

[8] Genuer, R., Poggi, J., and Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognition Letters 31*, 14 (2010), 2225–2236.

[9] Goodman, R., Speers, M., McLeroy, K., Fawcett, S., Kegler, M., Parker, E., Smith, S., Sterling, T., and N., W. Identifying and defining the dimensions of community capacity to provide a basis for measurement. *Health Education & Behavior 25*, 3 (1998), 258–278.

[10] Gutiérrez, N., Hilborn, R., and Defeo, O. Leadership, social capital and incentives promote successful fisheries. *Nature 470*, 7334 (2011), 386–389.

[11] Liberato, S., Brimblecombe, J., Ritchie, J., Ferguson, M., and Coveney, J. Measuring capacity building in communities: a review of the literature. *BMC public health 11*, 1 (2011), 850.

[12] Long, D., and Perkins, D. Community social and place predictors of sense of community: A multilevel and longitudinal analysis. *Journal of Community Psychology 35*, 5 (2007), 563–581.

[13] MacLellan-Wright, M., Anderson, D., Barber, S., Smith, N., Cantin, B., Felix, R., and Raine, K. The development of measures of community capacity for community-based funding programs in Canada. *Health Promotion International 22*, 4 (2007), 299–306.

[14] McMillan, D., and Chavis, D. Sense of community: A definition and theory. *Journal of community psychology 14*, 1 (1986), 6–23.

[15] Perkins, D., Brown, B., and Taylor, R. The ecology of empowerment: Predicting participation in community organizations. *Journal of Social Issues 52*, 1 (1996), 85–110.

[16] Press, M. Dimensions of community capacity building: A review of its implications in tourism development. *Journal of American Science 5*, 8 (2009), 172–180.

[17] Rupasingha, A., Goetz, S., and Freshwater, D. The production of social capital in US counties. *The journal of socio-economics 35*, 1 (2006), 83–101.

[18] Sengupta, N., Luyten, N., Greaves, L., Osborne, D., Robertson, A., Armstrong, G., and Sibley, C. Sense of community in New Zealand neighbourhoods: A multi-level model predicting social

capital. *New Zealand Journal of Psychology 42*, 1 (2013).

[19] Sheridan, J., and Tennison, J. Linking UK government data. In *LDOW* (2010).

[20] Sherrieb, K., Norris, F., and Galea, S. Measuring capacities for community resilience. *Social Indicators Research 99*, 2 (2010), 227–247.

[21] Simmons, A., Reynolds, R., and Swinburn, B. Defining community capacity building: is it possible? *Preventive medicine 52*, 3 (2011), 193–199.

[22] Siroky, D. Navigating random forests and related advances in algorithmic modeling. *Statistics Surveys 3* (2009), 147–163.

[23] Strobl, C., Malley, J., and Tutz, G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods 14*, 4 (2009), 323.

[24] Verikas, A., Gelzinis, A., and Bacauskiene, M. Mining data with random forests: A survey and results of new tests. *Pattern Recognition 44*, 2 (2011), 330–349.

[25] Xu, Q., Perkins, D. D., and Chow, J. C.-C. Sense of community, neighboring, and social capital as predictors of local political participation in China. *American journal of community psychology 45*, 3-4 (2010), 259–271.

APPENDIX

A. Data sources

Data publisher	Datasets collected	No. datasets	Datasets accessed through
Office of National Statistics	2011 Census datasets	17	Neighbourhood Statistics*, NOMIS**
Department for Communities and Local Government	English indices of deprivation 2010: income domain, living environment domain	2	data.gov.uk***
Department for Transport	Core Accessibility Indicators: Employment, Town Centres, Food Stores	3	gov.uk****
Home Office	Street-level crime, broken down by police force (whole of England)	1	data.police.uk*****
Total		23	

*http://neighbourhood.statistics.gov.uk/dissemination/.

**https://www.nomisweb.co.uk/.

***http://data.gov.uk/dataset/index-of-multiple-deprivation.

****https://www.gov.uk/government/collections/transport-connectivity-and-accessibility-of-key-services-statistics.

*****http://data.police.uk.

B. Sense of community variables, divided by categories

Socio-demographics

Indicator	**Datasets**	**Source**	**Variables selected**	**Variables included in the model**
Median age	KS102EW - Age structure (KS102EWDATA04)	2011 Census	KS102EW0019 - Median age	Same as selected.
Gender	QS104EW - Sex (QS104EWDATA04)	2011 Census	QS104EW0003 - Females, QS104EW0001 - Residents	Share of females over the general population (QS104EW003/QS104EW0001).
Length of residence in the UK	QS803EW - Length of residence in the UK (QS803EWDATA04)	2011 Census	QS803EW0002 - Born in the UK, QS803EW0003 - Resident in UK: Less than 2 years, QS803EW0004 - Resident in UK: 2 years or more but less than 5 years, QS803EW0005 - Resident in UK: 5 years or more but less than 10 years, QS803EW0006 - Resident in UK: 10 years or more	Same as selected.
Ethnic fragmentation	QS201EW - Ethnic group (QS201EWDATA04)	2011 Census	QS201EW0002 - White: English/Welsh/Scottish/Northern Irish/British, QS201EW0003 - White: Irish, QS201EW0004 - White: Gypsy or Irish Traveller, QS201EW0005 - White: Other White, QS201EW0006 - Mixed/multiple ethnic group: White and Black Caribbean, QS201EW0007 - Mixed/multiple ethnic group: White and Black African, QS201EW0008 - Mixed/multiple ethnic group: White and Asian, QS201EW0009 - Mixed/multiple ethnic group: Other Mixed, QS201EW0010 - Asian/Asian British: Indian, QS201EW0011 - Asian/Asian British: Pakistani, QS201EW0012 - Asian/Asian British: Bangladeshi, QS201EW0013 - Asian/Asian British: Chinese, QS201EW0014 - Asian/Asian British: Other Asian, QS201EW0015 - Black/African/Caribbean/Black British: African, QS201EW0016 - Black/African/Caribbean/Black British: Caribbean, QS201EW0017 - Black/African/Caribbean/Black British: Other Black, QS201EW0018 - Other ethnic group: Arab, QS201EW0019 - Other ethnic group: Any other ethnic group	Ethnic fragmentation (the selected variables were combined using the formula $ef = 1 - \sum_i (Race_i)^2$).
Religion	QS208EW - Religion (QS208EWDATA04)	2011 Census	QS208EW0002 - Christian, QS208EW0003 - Buddhist, QS208EW0004 - Hindu, QS208EW0005 - Jewish, QS208EW0006 - Muslim, QS208EW0007 - Sikh, QS208EW0008 - Other religion, QS208EW0009 - No religion	Same as selected.
Total	5		34	16

Socio-economic characteristics

Indicator	**Datasets**	**Source**	**Variables selected**	**Variables included in the model**
Employment sector	KS611EW - NS-SeC (KS611EWDATA04)	2011 Census	KS611EW0006 - 3. Intermediate occupations, KS611EW0011 - 8. Never worked and long-term unemployed	Same as selected.
Income	English indices of deprivation 2010: income domain (1871528)	Department for Communities and Local Government	Income Score	Same as selected.
People receiving benefits	Income support claimants	Department for Work and Pensions.	Total	Same as selected.
Level of qualification	QS501EW - Highest level of qualification (QS501EWDATA04)	2011 Census	QS501EW0006 - Level 3 qualifications, QS501EW0007 - Level 4 qualifications and above, QS501EW0008 - Other qualifications	Same as selected.
Total	4		7	7

Health

Indicator	Datasets	Source	Variables selected	Variables included in the model
Health conditions	QS302EW - General Health (QS302EWDATA04)	2011 Census	QS302EW0002 - Very good health, QS302EW0003 - Good health	Same as selected.
Total	1		2	2

Household composition

Indicator	Datasets	Source	Variables selected	Variables included in the model
Number of households with children	KS105EW - Household Composition (KS105EWDATA04)	2011 Census	KS105EW0006 - One Family Only; Married or Same-Sex Civil Partnership Couple; Dependent Children, KS105EW0009 - One Family Only; KS105EW0013 - Cohabiting Couple; Dependent Children; Other Household Types; With Dependent Children	Same as selected.
People married or living in a civil partnership	QS108EW - Living arrangement (QS108EWDATA04)	2011 Census	QS108EW0003 - Living in a Couple; Married (Persons) (Count), QS108EW0004 - Living in a couple: Cohabiting (opposite-sex), QS108EW0005 - Living in a couple: In a registered same-sex civil partnership or cohabiting (same-sex), QS108EW0006 - Not Living in a Couple; Total (Persons) (Count)	Same as selected.
Total	2		4	4

Tenure and housing category

Indicator	Datasets	Source	Variables selected	Variables included in the model
Homeowners and tenants	QS403EW - Tenure - People (QS403EWDATA04)	2011 Census	QS403EW0002 - Owned; Total, QS403EW0009 - Private Rented; Total	Same as selected.
Total	1		2	2

Social networks

Indicator	Datasets	Source	Variables selected	Variables included in the model
People providing unpaid care in the neighbourhood	QS301EW - Provision of unpaid care (QS301EWDATA04)	2011 Census	QS301EW0003 - Provides 1 to 19 hours unpaid care a week, QS301EW0004 - Provides 20 to 49 hours unpaid care a week, QS301EW0005 - Provides 50 or more hours unpaid care a week	Same as selected.
People working in the neighbourhood	Core Accessibility Indicator: employment centres (ACS041-2008)	Department for Transport	% AllCont PT/walk (EMPLO088) - % of users with access to employment centres by PT/walk	Same as selected.
Total	2		4	4

Resources and environment

Indicator	Datasets	Source	Variables selected	Variables included in the model
Town centres accessibility	Core Accessibility Indicator: town centres (ACS0408-2009)	Department for Transport	%AllCont PT/walk (TOWN064) - % of users with access to town centres within a reasonable time by PT/walk	Same as selected.
Commercial centres accessibility	Core Accessibility Indicator: food stores (ACS0407-2008)	Department for Transport	%AllCont PT/walk (SUPO064) - % of users with access to food stores within a reasonable time by PT/walk	Same as selected.
Religious organisations	QS420EW - Communal establishment management and type – Communal establishments (QS420EWDATA04)	2011 Census	QS420EW0032 - Other establishment: Religious	Same as selected.
Educational facilities	QS420EW - Communal establishment management and type – Communal establishments (QS420EWDATA04)	2011 Census	QS420EW0027 - Other establishment: Education	Same as selected.
Pollution	English indices of deprivation 2010: living environment domain (1871567)	Department for Communities and Local Government	Living environment score	Same as selected.
Property crimes	Street-level crime, broken down by police force (whole of England)	data.police.uk	Burglary, Vehicle crime (the variables were aggregated by local authority)	Aggregate number of burglaries per local authority; aggregate number of vehicle crime per local authority.
Crimes against the person	Street-level crime, broken down by police force (whole of England)	data.police.uk	Criminal damage, Violent crime (the variables were aggregated by local authority)	Aggregate number of criminal damages per local authority; aggregate number of violent crimes per local authority.
Total	5		7	9

Sense of community measure - dependent variable

Dataset	Source	Variable used	Description*
National Indicator Set	Department for Communities and Local Government (accessed through data.gov.uk**)	NI 002 - Percentage of people who feel that they belong to their neighbourhood	The proportion of the adult population who feel ‘fairly strongly’, or ‘very strongly’ that they belong to their immediate neighbourhood. Belonging: Respondents will be said to feel they belong to their area if they say they feel they belong “very strongly” or “fairly strongly”.

*As provided within the dataset.

**http://data.gov.uk/dataset/ni-002-percentage-of-people-who-feel-that-they-belong-to-their-neighbourhood.

Total datasets collected	**Total variables selected**	**Total variables included in the model**
20	65	48 (+ 1 dependent variable)

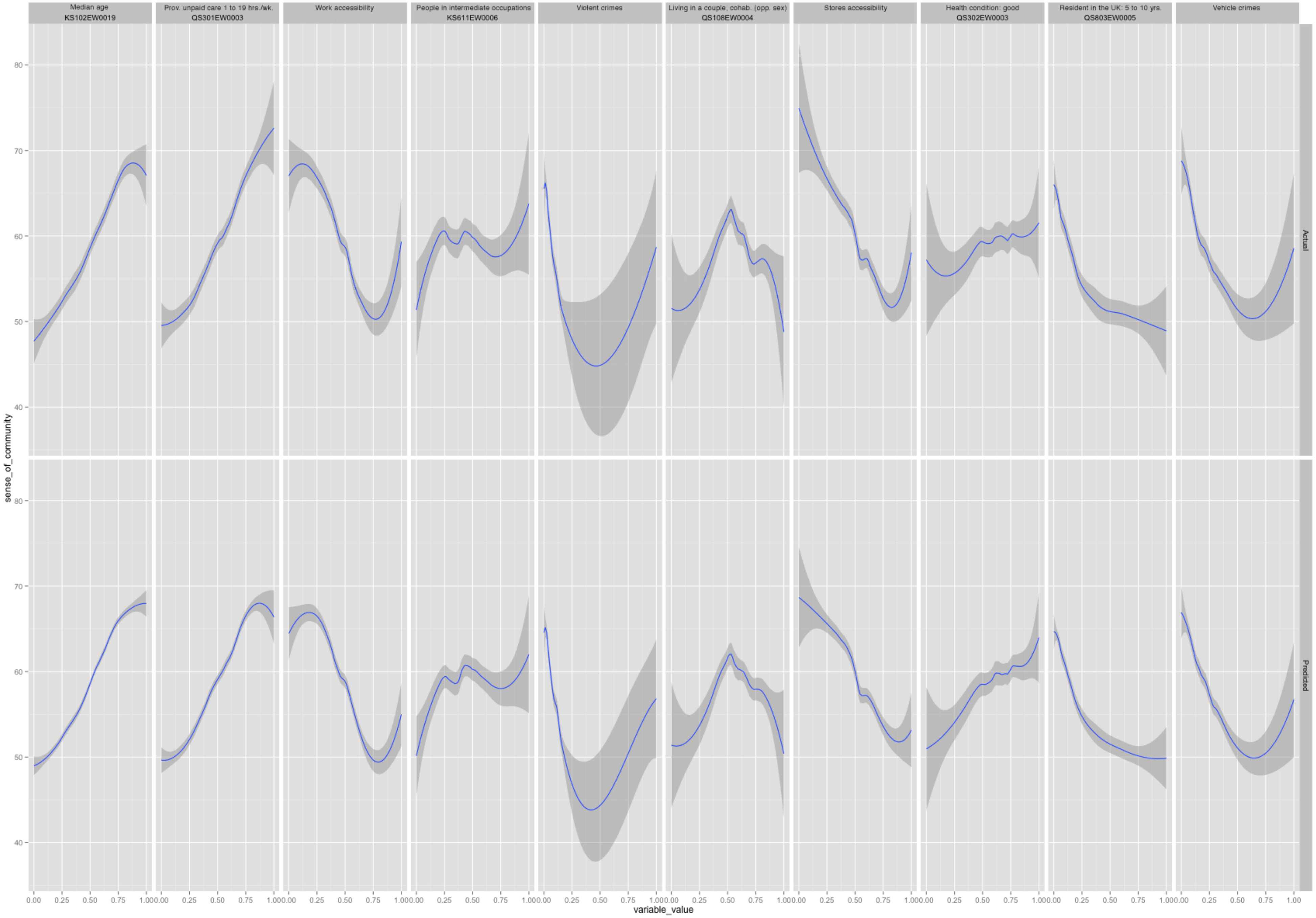


Figure B.1: Sense of community values – actual (above) and predicted (below) – plotted against the ten most predictive variables in the dataset (other variables were left out due to their low predictivity). The gray area represents the range of the actual values, whereas the blue line is their average. Variables are normalised to 0-1.

C. Participation variables, divided by categories

Socio-demographics

Indicator	**Datasets**	**Source**	**Variables selected**	**Variables included in the model**
Median age	KS102EW - Age structure (KS102EWDATA04)	2011 Census	KS102EW0019 - Median age	Same as selected.
Gender	QS104EW - Sex (QS104EWDATA04)	2011 Census	QS104EW0003 - Females, QS104EW0001 - Residents	Share of females over the general population (QS104EW003/QS104EW0001).
Length of residence in the UK	QS803EW - Length of residence in the UK (QS803EWDATA04)	2011 Census	QS803EW0002 - Born in the UK, QS803EW0003 - Resident in UK: Less than 2 years, QS803EW0004 - Resident in UK: 2 years or more but less than 5 years, QS803EW0005 - Resident in UK: 5 years or more but less than 10 years, QS803EW0006 - Resident in UK: 10 years or more	Same as selected.
Ethnic fragmentation	QS201EW - Ethnic group	2011 Census	QS201EW0002 - White: English/Welsh/Scottish/Northern Irish/British, QS201EW0003 - White: Irish, QS201EW0004 - White: Gypsy or Irish Traveller, QS201EW0005 - White: Other White, QS201EW0006 - Mixed/multiple ethnic group: White and Black Caribbean, QS201EW0007 - Mixed/multiple ethnic group: White and Black African, QS201EW0008 - Mixed/multiple ethnic group: White and Asian, QS201EW0009 - Mixed/multiple ethnic group: Other Mixed, QS201EW0010 - Asian/Asian British: Indian, QS201EW0011 - Asian/Asian British: Pakistani, QS201EW0012 - Asian/Asian British: Bangladeshi, QS201EW0013 - Asian/Asian British: Chinese, QS201EW0014 - Asian/Asian British: Other Asian, QS201EW0015 - Black/African/Caribbean/Black British: African, QS201EW0016 - Black/African/Caribbean/Black British: Caribbean, QS201EW0017 - Black/African/Caribbean/Black British: Other Black, QS201EW0018 - Other ethnic group: Arab, QS201EW0019 - Other ethnic group: Any other ethnic group	Ethnic fragmentation (the selected variables were combined using the formula $ef = 1 - \sum_i (Race_i)^2$).
Proficiency in English	QS205EW - Proficiency in English (QS205EWDATA04)	2011 Census	QS205EW0005 - Main language is not English (English or Welsh in Wales): Cannot speak English well, QS205EW0006 - Main language is not English (English or Welsh in Wales): Cannot speak English	Same as selected.
Total	5		28	10

Socio-economic characteristics

Indicator	Datasets	Source	Variables selected	Variables included in the model
Employment status	QS601EW - Economic activity (QS601EWDATA04)	2011 Census	QS601EW0009 - Economically active: Unemployed	Same as selected.
Women in employment	KS603EW - Economic Activity - Females (KS603EWDATA04)	2011 Census	KS603EW0002 - Economically Active; Employee; Part-Time, KS603EW0003 - Economically Active; Employee; Full-Time.	Same as selected.
Hours worked	QS604EW - Hours worked (QS604EWDATA04)	2011 Census	QS604EW0006 - Full-time: 31 to 48 hours worked, QS604EW0007 - Full-time: 49 or more hours worked	Same as selected.
Income	English indices of deprivation 2010: income domain (1871528)	Department for Communities and Local Government	Income score	Same as selected.
People receiving benefits	Income support claimants	Department for Work and Pensions	Total	Same as selected
Socio-economic status	KS611EW - NS-SeC (KS611EWDATA04)	2011 Census	KS611EW0002 - 1. Higher managerial, administrative and professional occupations, KS611EW0005 - 2. Lower managerial, administrative and professional occupations, KS611EW0006 - 3. Intermediate occupations, KS611EW0007 - 4. Small employers and own account workers, KS611EW0008 - 5. Lower supervisory and technical occupations, KS611EW0009 - 6. Semi-routine occupations, KS611EW0010 - 7. Routine occupations, KS611EW0011 - 8. Never worked and long-term unemployed, KS611EW0015 - L15 Full-time students	Same as selected.
Level of qualification	QS501EW - Highest level of qualification (QS501EWDATA04)	2011 Census	QS501EW0006 - Level 3 qualifications, QS501EW0007 - Level 4 qualifications and above, QS501EW0008 - Other qualifications	Same as selected.
Total	7		19	19

Health

Indicator	Datasets	Source	Variables selected	Variables included in the model
Health conditions	QS302EW - General Health (QS302EWDATA04)	2011 Census	QS302EW0002 - Very good health, QS302EW0003 - Good health	Same as selected.
Total	1		2	2

Household composition

Indicator	Datasets	Source	Variables selected	Variables included in the model
Number of households with children	KS105EW - Household Composition (KS105EWDATA04)	2011 Census	KS105EW0006 - One Family Only; Married or Same-Sex Civil Partnership Couple; Dependent Children, KS105EW0009 - One Family Only; KS105EW0013 - Cohabiting Couple; Dependent Children; Other Household Types; With Dependent Children	Same as selected.
People married or living in a civil partnership	QS108EW - Living arrangement (QS108EWDATA04)	2011 Census	QS108EW0003 - Living in a Couple; Married (Persons) (Count), QS108EW0004 - Living in a couple: Cohabiting (opposite-sex), QS108EW0005 - Living in a couple: In a registered same-sex civil partnership or cohabiting (same-sex), QS108EW0006 - Not Living in a Couple; Total (Persons) (Count)	Same as selected.
Total	2		7	7

Tenure and housing category

Indicator	Datasets	Source	Variables selected	Variables included in the model
Homeowners, tenants and living in a social rented house	QS403EW - Tenure - People (QS403EWDATA04)	2011 Census	QS403EW0002 - Owned; Total, QS403EW0006 - Private Rented; Total, QS403EW0009 - Social Rented; Total	Same as selected.
Total	1		3	3

Social networks

Indicator	Datasets	Source	Variables selected	Variables included in the model
People providing unpaid care in the neighbourhood	QS301EW - Provision of unpaid care (QS301EWDATA04)	2011 Census	QS301EW0003 - Provides 1 to 19 hours unpaid care a week, QS301EW0004 - Provides 20 to 49 hours unpaid care a week, QS301EW0005 - Provides 50 or more hours unpaid care a week	Same as selected.
People working in the neighbourhood	Core Accessibility Indicator: employment centres (ACS041-2008)	Department for Transport	% AllCont PT/walk (EMPLO088) - % of users with access to employment centres by PT/walk	Same as selected.
Total	2		4	4

Resources and environment

Indicator	Datasets	Source	Variables selected	Variables included in the model
Religious organisations	QS420EW - Communal establishment management and type – Communal establishments (QS420EWDATA04)	2011 Census	QS420EW0032 - Other establishment: Religious	Same as selected.
Professional organisations	QS420EW - Communal establishment management and type – Communal establishments (QS420EWDATA04)	2011 Census	QS420EW0012 - Medical and care establishment: Registered Social Landlord/Housing Association: Total	Same as selected.
Educational facilities	QS420EW - Communal establishment management and type – Communal establishments (QS420EWDATA04)	2011 Census	QS420EW0027 - Other establishment: Education	Same as selected.
Total	1		3	3

Sense of community measure - dependent variable

Dataset	Source	Variable used	Description[*]
National Indicator Set	Department for Communities and Local Government (accessed through data.gov.uk[**])	NI 003 - Civic participation in the local area	Civic participation is one of the principal means by which individuals exercise their empowerment for the benefit of the locality, often at the same time increasing their own level of empowerment. Contributing to a decision-making group requires a degree of personal confidence combined with a willingness to be a conduct for wishes and needs of other residents. An increase in the number and diversity of people taking on such roles can help to create fairer, more inclusive policies whilst spreading the perception that public decision making is accessible to the influence of all legitimate interests.

[*] As provided within the dataset.

[**] http://data.gov.uk/dataset/ni-003-civic-participation-in-the-local-area.

Total datasets collected	Total variables selected	Total variables included in the model
19	66	48 (+ 1 dependent variable)

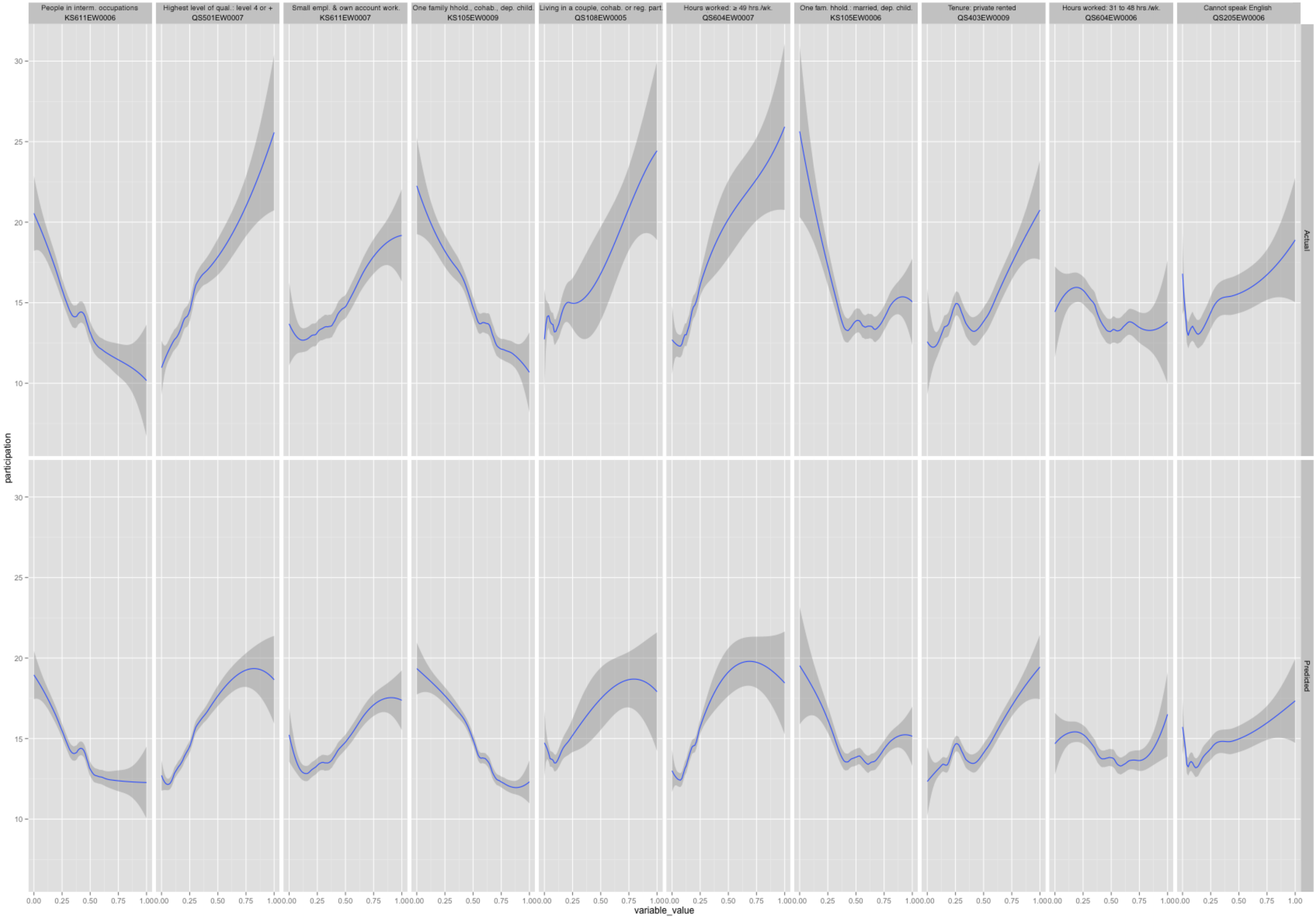


Figure C.1: Participation values – actual (above) and predicted (below) – plotted against the ten most predictive variables in the dataset (other variables were left out due to their low predictivity). The gray area represents the range of the actual values, whereas the blue line is their average. Variables are normalised to 0-1.