Policy and Internet

# Predicting Sense of Community and Participation by Applying Machine Learning to Open Government Data

**Alessandro Piscopo, Ronald Siebes, and Lynda Hardman**

*Community capacity is used to monitor socioeconomic development. It is composed of a number of dimensions that can be measured to understand issues possibly arising in the implementation of a policy or of a project targeting a community. Measuring these dimensions is thus highly valuable for policymakers and local administrator, though expensive and time consuming. To address this issue, we evaluated their estimation through a machine learning technique—Random Forests—applied to secondary open government data and determined the most important variables for prediction. We focused on two dimensions: sense of community and participation. The variables included in the data sets used to train the predictive models complied with two criteria: nationwide availability and sufficiently fine-grained geographic breakdown, that is, neighborhood level. Our resultant models are more accurate than others based on traditional statistics found in the literature, showing the feasibility of the approach. The most determinant variables in our models were only partially in agreement with the most influential factors for sense of community and participation according to the social science literature consulted, providing a starting point for future investigation under a social science perspective. Moreover, due to the lack of geographic detail of the outcome measures available, further research is required to apply the predictive models to a neighborhood level.*

**KEY WORDS:** open data, machine learning, e-Government, sense of community, civic participation

## Introduction

Community-based approaches are widely employed in publicly or privately funded programs targeted to promoting socioeconomic development. Building the capacity of a community, or community capacity (CC), is key for these approaches, 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 (Liberato, Brimblecombe, Ritchie, Ferguson, & Coveney, 2011). It is used mainly in the implementation of public health policies, with applications in other fields (Press, 2009), for example, tourism.

All the definitions of CC agree that this is composed of several dimensions (Simmons, Reynolds, & Swinburn, 2011). Those included more often are learning



1944-2866 
Published by Wiley Periodicals, Inc., 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford, OX4 2DQ.

opportunities and skills development; resource mobilization; leadership; participatory decision making (or participation); asset-based approach; sense of community; communication; partnership/linkages/networking; and development pathway (Liberato et al., 2011). High levels of CC increase the possibility of policies targeting a community to be successful (Goodman et al., 1998; Simmons et al., 2011); since this is affected by any change in its dimensions, it is important to understand which are deficient and which initiatives should be taken to improve them (Simmons et al., 2011). Furthermore, their evaluation facilitates policymakers 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. To gauge CC, local surveys are usually organized (MacLellan-Wright et al., 2007), but they may not be feasible due to their high costs. This also hampers the realization 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 development can be associated with health outcomes" (Liberato et al., 2011, p. 850). The absence of such measurements is reflected in a greater focus in the literature on describing the process of CC building, rather than on its measurement (Liberato et al., 2011).

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, when organizing a local survey is not possible. In addition, it would raise the self-assessment ability of communities themselves, improve the accountability of local administrations, be an instrument to perform a longitudinal study of CC on a larger scale.

An alternative approach to obtain CC dimension measures, investigated in this research, applies predictive algorithms to secondary data, that is, data collected primarily for other purposes, related to topics other than social dimensions, such as demographics or socioeconomic 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 investigation focused on two CC dimensions: sense of community and participation.

The main question we pose is to what extent can we predict measures of participation and sense of community through applying machine learning to secondary data? In order to be possible to use them in the context chosen, measures have to comply with two criteria: consistent nationwide applicability, meaning that the measure has to be available for any area within the context of our study; and high geographic precision, that is, they must be detailed at neighborhood level (Chainey, 2008). A secondary research question is to determine which variables have the highest influence for predicting sense of

community and participation in the context chosen and whether they are in agreement with those determined using other models in the literature.

The primary contribution of this research lies in demonstrating the feasibility of a machine-learning based approach to obtain measures of social dimensions for local communities with few requirements in terms of economic and time resources. In particular, we show that Random Forests is suitable for this purpose, as it is accurate, able to deal with small data sets and nonlinear data, and provides information about how each variable in the data set contributes to prediction accuracy. Finally, our study identifies a selection of openly available data sets which are relevant for predicting sense of community and participation.

The paper is structured as follows. Work relevant to the choice of social dimensions indicators and of the predictive algorithm is presented in the Related Work section.. The method used for selecting the relevant variables for the models, as well as the data gathering and processing, is described in the Methods section. 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 the Data Description section. Finally, the last three sections are dedicated to presenting, discussing, and drawing conclusions from the results of this study.

## Related Work

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

### *Social Dimension Indicators*

The first step to build our predictive models was to select the variables to include. Although our aim was to predict CC dimensions using secondary data, we needed to identify the data sets relevant for each dimension studied. Predictive models of social dimensions are generally built on a selected number of metrics, which are low-level measures of high-level relevant indicators (or predictors) (Long & Perkins, 2007; MacLellan-Wright et al., 2007; Sengupta et al., 2013; Sherrieb, Norris, & Galea, 2010). For example, the percentage of homeowners in a neighborhood can be used as a metric of an indicator such as neighborhood residents' type of tenure. Beyond predicting the level of a social dimension, these models aim at describing the type of relationship existing between that and the indicators used (Long & Perkins, 2007; Sengupta et al., 2013), or at building an index to measure a concept, by using proxy (secondary) data (Sherrieb et al., 2010). To select the indicators, their relevance can be assessed on the basis of theoretical assumptions (Dekker, 2007; Long & Perkins, 2007), confirmed or contradicted through an analysis of the data collected—often in a survey organized specifically for the study. Another approach is the submission of several indicators derived from a literature review to a group of experts, who assess the suitability of those for a determined context (MacLellan-Wright et al.,

2007). It was out of our scope to build models explaining which factors influenced the social dimensions chosen and the choice of a group of experts would have been against our aim to provide fast and inexpensive measures of CC dimensions. Whereas the aforementioned selections included direct measurements of social dimensions, which were collected using surveys made on population samples, we wanted to use only secondary data, such as demographics and socioeconomic data, collected on the overall population. Nevertheless, these studies made use of a sound selection approach to be followed for choosing the appropriate metrics for each indicator, and were a reliable source to identify indicators for participation and sense of community. For each indicator found in the literature, we compiled a "wishlist" of metrics that possibly described it, in a top–down fashion. The wishlist was followed to find relevant variables within a number of data sources and data sets. We followed the process in Sherrieb et al. (2010), who create an index for community resilience, with the exception that we did not further reduce the indicators according to their intercorrelation.

In addition, Venerandi, Quattrone, Capra, Quercia, and Saez-Trumper (2015) attempt to predict measures of urban deprivation using secondary data. The authors utilize data from Foursquare and OpenStreetMap to compare with measures of deprivation for neighborhoods in large- and mid-size English cities, by analyzing correlations between a number of features in their data sets and the sought measure. They showed correlation at multiple features level and predicted the independently assessed deprivation measures with an accuracy comparable to the state of the art. Furthermore, they analyzed the highest correlated variables to determine nine themes that could be relevant for urban deprivation. However, as the authors themselves point out (Mashhadi, Quattrone, & Capra, 2013; Venerandi et al., 2015), the data sets employed did not provide uniform coverage throughout the country, namely one of the requirements for our measures, and were socially biased, that is, were produced by generally young, educated, and wealthy users. Finally, social media data, that is, Twitter data, are used by Quercia, Seaghdha, and Crowcroft (2012) to predict indices of multiple deprivation, through the application of linear regression to topics modeled on tweets' texts. Also in this case, the models built are able to partially explain the variability of the outcome measure, but they are affected by the social and geographic bias deriving from the characterization of Twitter users.

### *Prediction Techniques*

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

Because of the strong assumptions formulated on the structure underlying the data (Breiman, 2001), 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—that is, a

**Table 1.** Main Differences Between Standard Statistics and Data Mining (Breiman, 2001; Friedman, 1998)

| | Standard Statistics | Data Mining |
|---|---|---|
| Example techniques | Linear regression, factor analysis, ANOVA | Neural networks, decision trees, SVMs |
| Domain knowledge | Based on strong theoretical assumptions | Relies 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 | Goodness-of-fit tests, residual examination | Prediction accuracy |

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 required 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 (Friedman, 1998). 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 (Berk, 2006). This is in accordance with our purpose of building an easily deployable system to predict measures of CC dimensions.

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, predictions made by support vector machines (SVMs), one of the most accurate algorithms (Verikas, Gelzinis, & Bacauskiene, 2011), are difficult to explain (Barbella et al., 2009). 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 (Strobl, Malley, & Tutz, 2009b), while providing high prediction accuracy, compared with other algorithms (Verikas et al., 2011). This technique, applied already to several fields, such as genetics, psychology, and organization management (Gutiérrez, Hilborn, & Defeo, 2011), is suitable for predicting either categorical (classification) or continuous outcome values (regression tasks). The characteristics of Random Forests, which grows successive decision trees, for each one using a random sample of the training data, make it robust to overfitting (Siroky, 2009) and avoid the problems derived by the "multiplicity of good models" (Breiman, 2001, p. 200). This definition refers to the possibility of building a high number of equally predictive models in the presence of high-dimensional data sets, by removing even small subsets (2–3 percent) (Breiman, 2001). 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 $n \ll p$ (Strobl et al., 2009b). Another advantage of Random Forests is 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 (Strobl et al., 2009b). It computes the contribution of a variable for prediction accuracy by randomly permuting it, evaluating the model 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" (Strobl et al., 2009b, p. 324).

Therefore, several reasons made Random Forests suitable for our purposes: high prediction accuracy and interpretability of results, which were suitable to create a predictive model for use in real settings and to obtain insights on the most relevant variables for this task; robustness to overfitting and to the multiplicity of good models problems; suitability for data sets with many variables and few instances, which were appropriate for the data sets created, as these had a large $p$ (about 50 variables) and small $n$ (about 300 instances).

## Methods

After explaining the selection criteria for CC dimensions, we illustrate how we picked, collected, and processed the data. The choice of CC dimensions and the data selection proceeded in parallel, so their outcomes influenced one another.

### *Selection of CC Dimensions*

We investigated only two CC dimensions, sense of community and participation, on the basis of the availability of measures to be used as dependent variables for training our predictive models. The available measures had to satisfy three requirements, on the basis of our aims of creating a fast, inexpensive, and wide applicable method to gauge CC:

- To match as closely as possible the social dimensions we wanted to investigate. 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.
- To have a consistent national coverage.
- To be able to provide information about the geographic detail of a small to medium-sized neighborhood (up to a few thousand residents). Smaller areas also provide the advantage of increasing the number of instances for the data set used to train our model, as each measurement represents an instance in the data set. Using the nomenclature of the U.K. Office of National Statistics (ONS) geography, employed with minor changes in the 2001 and 2011 Censuses, the best geographic breakdown for this was the Lower Super Output Area (LSOA). Its level of detail describes appropriately a neighborhood, while providing wider availability of data sets than the smaller ONS statistical subdivision, the Output Area (OA; see Table 2 and Figure 1).

**Table 2.** Office of National Statistics England and Wales Geography and Local Authorities Statistics (2011)

| 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 | – |
| LA | 162,615 | 75,188 | 325 | – |

*Source*: ons.gov.uk.
*Notes*: Characteristics of Scotland Geography slightly differ. LSOA is the level that provides the best combination of level of detail and availability of data sets.

Notwithstanding the wide availability of national surveys investigating social dimensions in the United Kingdom, none of them satisfied all three of these requirements. According to the descriptions provided, which were somewhat inexhaustive, the National Indicators NI 002 and NI 003 are the measures that represent sense of community and participation more closely, have national coverage, and the most detailed geographic level available.

NI 002[2] was chosen as a measure for sense of community, whereas NI 003[3] was used for participation; their geographic breakdown is the local authority (LA) level.

*Sense of community and participation measures.* 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" (McMillan & Chavis, 1986, p. 9). It plays an essential role in CC building, as it increases the active membership—which is at the basis of participation—influences the collective norms and values, and improves the mobilization of resources (Goodman et al., 1998). The measure used to train our model for sense of community was NI 002 (*percent of people who feel that they belong*

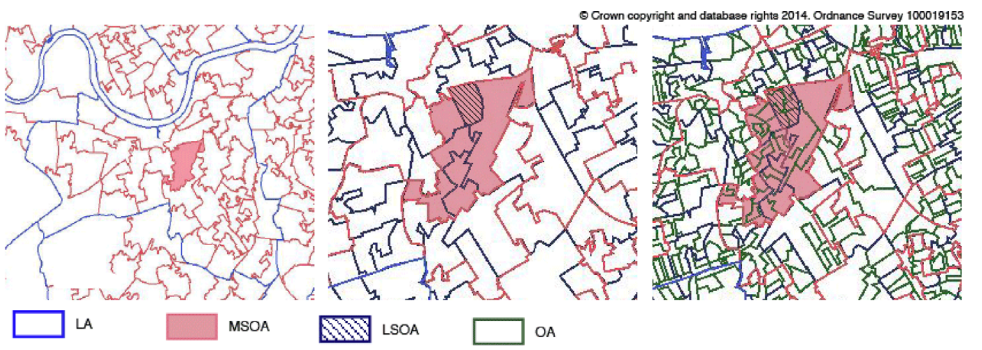



**Figure 1.** Relative Sizes of Output Areas (OA), Super Output Areas (LSOA, MSOA), and Local Authorities (LA). *Note*: Larger areas are aggregations of smaller ones.

*to their neighborhood*), which is constructed on the basis of the responses to the question "How strongly do you feel you belong to your immediate neighborhood?" 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 entirely the complexity of sense of community, we found NI 002 a sufficiently accurate measure for it and the best available.

Participation is defined as "people's engagement in activities within the community" (Sharifuddin et al., 2015, p. 3670). It is an essential quality of CC, as community members may gain an understanding of and act on issues concerning the community as a whole only by participating in small groups or smaller organizations (Goodman et al., 1998). 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 (Goodman et al., 1998). NI 003 (*Civic participation in the local area*) provided an appropriate measure for this social dimension. It 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. Therefore, a higher rate of positive answers indicates a higher civic participation within the community.

The geographic breakdown of NI 002 and NI 003 is the 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). They are constructed on responses collected within the 2008 Place Survey, now discontinued. This survey was administered by LAs and "provides information on people's perceptions of their local area and the local services they receive."[4] It used a multistage 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 measures provide continuous values, with higher ones indicating better performance, that is, higher levels of sense of community or participation.

## *Data Gathering and Processing*

*Data Selection Criteria.* We selected variables for our models on the basis of the relevant indicators of participation and sense of community. For each indicator, we formulated a hypothesis about which metrics were the most appropriate to describe it (Sherrieb et al., 2010). By following this process for each indicator in our selection, we compiled a wishlist of variables to be included. Subsequently, we checked which variables in the data sets available from English open government data sources matched the ones in our wishlist: the matching ones were included in our data set. Variables not present in our wishlist but clearly related to the indicators selected were included as well. As an example of the process followed, Dekker (2007) states that social networks within a neighborhood may be a relevant indicator of participation; following the variables used in this study, which relate to another country though, we first built a wishlist of

measures. These included, among others, the presence of family and friends in the neighborhood and the frequency of relations with the neighbors, which were either not available from the data sources selected, or did not meet the requirements previously set. Measures present in the data sources available, specifically the number of people providing unpaid care and the percentage of people working in the neighborhood, provided an indirect measurement of social networks, therefore, we included them in our models.

To be suitable for selection, data sets had to comply with three criteria: geographical coverage, geographical detail, and time (see Table 3). 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; and they had to be available at the geographic breakdown suitable for small neighborhoods. In order to make our selection theoretically suitable for smaller areas, this latter condition was followed for data sets to be included and they were available also at LSOA level, which provided the best combination of geographical detail and availability. However, the sense of community and participation measures used had England-wide coverage at LA level, therefore, the dependent variables selected had the same characteristics. The time criterion required that data should be available for a time span as close as possible to the dependent variables. NI 002 and NI 003 referred to 2008, but we included data up to 2011, the year of the last U.K. census on the general population and the closest year for which crime data were available. This was not an issue, considering the slow evolution times of social dimensions (Sherrieb et al., 2010). Finally, indicators for which no measures were available were discarded.

*Data Cleaning and Preparation.* The data sets collected complied with the quality standards of the ONS and other government departments, that is, accuracy, coherence, and comparability, therefore, contained no missing values or rogue attributes. The variables depending on LA size were normalized, dividing them by the total number of units to which they referred, for example, number of residents or number of households. Data related to the ethnic composition of the population were used to calculate ethnic fragmentation, which is correlated with participation and social cohesion (Alesina & La Ferrara, 2000).

*Data processing.* We aim at building models to predict levels of sense of community and participation. Both these presented continuous values—the

**Table 3.** Data Selection Criteria

| Criterion (Condition Sought) | Condition Available |
|---|---|
| Geographical coverage (nationwide coverage) | England |
| Geographical detail (neighborhood level) | LA |
| Time (closeness to social dimension measures used) | 2008 (dependent variables)–2011 |

*Note*: Data used for the prediction models were determined by the characteristics of the social measures available.

outcome variables—therefore, Random Forests was applied on the selected variables to solve 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. In other words, each decision tree is built using a random subset of instances and evaluated using the ones that were not included in this subset. The accuracy of each observation is calculated only taking into account the trees in which it was not comprised. Finally, the performance of the whole model is calculated by averaging the results of all the trees. This feature was particularly valuable in view of the small number of instances in our data sets, as it allows not to use separate training and test sets. Furthermore, OOB-based estimates are considered to be more realistic and conservative than the ones resulting from the application of a model to a new test set (Strobl et al., 2009b). The Random Forests algorithm was applied through its *R* implementation in the package *party*, whose importance measures have shown to be reliable also in case of highly correlated variables (Strobl, Hothorn, & Zeileis, 2009a).

We tuned the algorithm used by setting three parameters, *mtry*, that is, the number of variables randomly chosen at each split; *ntree*, that is, the number of trees in the forest; and *nodesize*,[5] which indicates the minimal number of instances in the terminal nodes of each tree (Statnikov, Wang, & Aliferis, 2008). An optimal setting of these parameters may provide a higher and more reliable prediction accuracy, with a more stable model (Genuer, Poggi, & Tuleau-Malot, 2010; Strobl et al., 2009a). Appropriately tuned values of *mtry* and *ntree* ensure a lower bias in selecting important variables (Verikas et al., 2011). Moreover, models with higher *mtry* have been found to better convey conditional importance in presence of highly correlated variables (Strobl et al., 2009b). To tune each model, we first set *ntree*, *mtry*, and *nodesize* to their default values for regression ($ntree = 500$, $mtry = p/3$, $nodesize = 1$). Afterward, we increased these parameters by 100 (*ntree*) and by 1 (*mtry* and *nodesize*), and tested prediction accuracy of each combination of these values, until no improvements were observed. The accuracy measures were mean squared error (MSE) and $R^2$, 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 $R^2 = 1 - \frac{SS_E}{SS_T}$, 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*, *ntree*, and *nodesize* were evaluated by comparing them to predictive models of social dimensions found in the literature.

The variable importance was computed accounting for the conditional importance of the variables. This value measures how much each variable contributes alone to prediction accuracy. Relative rankings of variables were used to assess the results relative to their predictivity. We did not report the importance values produced by the algorithm, as these are not comparable among different studies (Strobl et al., 2009b). However, in order to better convey the degree of predictivity of each variable with respect to the others, we provide the ratios among their importance values.

Finally, we applied a heuristic from Strobl et al. (2009b) to identify variables irrelevant for prediction: Variables can be considered informative and important if their importance score is above the absolute value of the variable with the lowest negative score. At the basis of that is the assumption that irrelevant and uninformative variables present importance values randomly varying around zero.

## Data Description

A total of 23 data sets were collected, 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. 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.[6] QS provides the maximum possible detail (OA), whereas KS is often available only for LSOAs and MSOAs. The indicators selected covered various areas, such as socioeconomic characteristics, socio-demographics, and housing conditions. The data sets related to sense of community and participation are shown in Tables 4 and 5. Both the data sets created for our predictive models had 316 instances, each corresponding to an English LA. The difference between the number of values of NI 002 and NI 003 and the final number of instances in the data sets was due to divergences between the administrative geographies used in some data sets and to modifications to the number of LAs between 2008 and 2011. Therefore, not all of the English LAs were included in the data sets. This is a summary of the characteristics of the variables included in each data set:

- The sense of community data set had 48 continuous independent variables and one continuous dependent variable (NI 002) (see Figure 2). This had a maximum value of 75.1 and a minimum one of 42.8.
- The participation data set had 48 continuous independent variables—the two data sets had the same number of variables by coincidence—and one continuous dependent variable (NI 003) (see Figure 3; the equivalence of the number of variables in the two data sets is accidental). This had a maximum value of 25.7 and a minimum of 7.6.

## Results

### *Sense of Community*

The optimal settings for the sense of community model were *mtry* 44, *ntree* 1,200, and *nodesize* 7. Using these values, the model yielded an MSE of 9.5 and an $R^2$ of 76.6 percent (Figure 4). The prediction accuracy did not increase by growing further trees, raising the number of variables chosen at each split, or varying the minimum size of terminal nodes. According to the heuristic enunciated in the Data Processing subsection, only 7 variables out of 48 could be regarded as not important for prediction. The median age of the population was the most predictive variable,

**Table 4.** Indicators and Data Sets Collected for Sense of Community

| Category | Indicators | No. of Data Sets (Year) | Source Data Sets | No. of Variables Used[a] |
|---|---|---|---|---|
| Socio-demographics | Gender, age, ethnic composition of neighborhood, religion | 5 (2011) | 2011 Census | 34 (16) |
| Socioeconomic characteristics | Employment sector, employment status, income level, 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 | Marital status, number of children | 2 (2011) | 2011 Census | 7 (7) |
| Tenure and housing category | Tenure type | 1 (2011) | 2011 Census | 3 (3) |
| Social networks | Activities and relationships in the neighborhood | 2 (2011) | 2011 Census, core accessibility indicators | 4 (4) 7 (9) |
| Resources and environment | Religious organizations, educational facilities, commercial facilities, pollution, accessibility, crime levels. 2011 Census, core accessibility indicators, English indices of deprivation 2010, data.police.uk | 4 (2011), 1 (2010), 1 (2009) | | |
| Total | | 20 | | 66 (48) |

[a]In parentheses, the number of variables included in the model after aggregation.

followed by the share of people providing 1–19 hours unpaid care a week (ratio between its importance value and the higher ranking variable one: 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. The relative importance of variables is shown in Figure 2.

### *Participation*

The optimal settings for the participation model were *mtry* 38, *ntree* 1,000, and *nodesize* 4 which yielded MSE 3.7 and $R^2$ 62.6 percent (Figure 5). Growing further trees, increasing the number of variables at each split, or setting different node size values did not improve the accuracy of the model. According to the heuristic enunciated in the Data Processing subsection, only 10 variables out of 48 were neither informative nor important. The variable with the highest importance value was the proportion of people in intermediate occupations, followed by the proportion of people with a level $\geq$4 of education (ratio between its

**Table 5.** Indicators and Data Sets Collected for Participation

| Category | Indicators | No. of Data Sets (Year) | Source Data Sets | No. of Variables Used[a] |
|---|---|---|---|---|
| Socio-demographics | Gender, age, ethnic composition of neighborhood, proficiency in English | 6 (2011) | 2011 Census | 28 (10) |
| Socioeconomic characteristics | Employment status, women in employment, income level, socioeconomic status, level of qualification | 7 (2011), 1 (2010) | 2011 Census, English indices of deprivation 2010, benefits claimants | 19 (19) |
| Health | Health conditions | 1 (2011) | 2011Census | 2 (2) |
| Households composition | Marital status, number of children | 2 (2011) | 2011 Census | 7 (7) |
| Tenure and housing category | Tenure type | 1 (2011) | 2011 Census | 3 (3) |
| Social networks | Activities and relationships in the neighborhood | 1 (2011), 1 (2009) | | |
| | 2011 Census, core accessibility indicators | 4 (4) | | |
| Resources and environment | Religious organizations, professional organizations, education facilities | 1 (2011) | 2011 Census | 3 (3) |
| Total | | 21 | | 66 (48) |

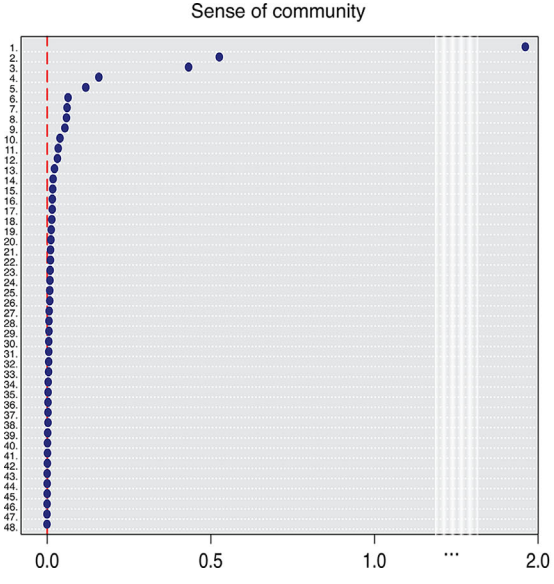
[a]In parentheses, the number of variables included in the model after aggregation.

importance value and the higher ranking variable one: 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) (see Figure 3).

## Discussion

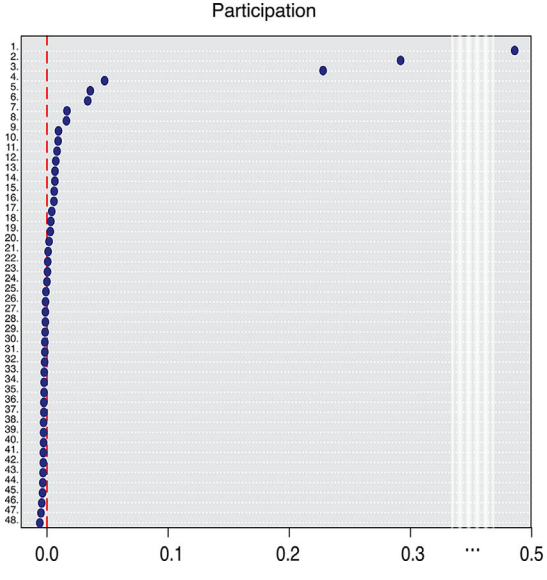
### *Accuracy of the Model and Applicability*

The sense of community model obtained the best results for explaining the variation of the dependent variable (see Figures 4 and 5). The higher MSE for this model can be related to the higher range of the sense of community measure. Neither of the models was suitable to predict CC dimensions at neighborhood 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 method investigated could be a valuable resource for local administrators and policymakers, who could take advantage of them to obtain estimations of the social characteristics of their communities. The characteristics could include not only sense of community and




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 2.** Variable Importance for Sense of Community. *Notes*: The dashed line indicates zero; the names in parentheses indicate the source data set 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 3.** Variable Importance for Participation. *Notes*: The dashed line indicates zero; the names in parentheses indicate the source data set 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.

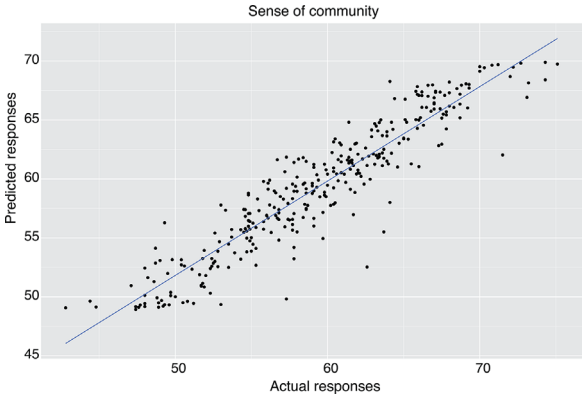



**Figure 4.** Sense of Community (NI 002): Plot of the Predicted Responses to the Actual Ones. *Note*: The closer the predicted responses are to the line, the better the model fits the actual data (different scale from participation, Figure 5).

participation, but may extend to others, provided that ground truth measures are made available. These could be obtained by means of purposely organized surveys, which could be done at regular intervals, whereas between them a machine learning approach would provide inexpensive, still accurate measures.

The prediction accuracy was high, compared to previous studies in which parametric models were used. To the best of our knowledge, no similar experiments to predict sense of community and participation have been carried out in the same context of this research, therefore, we relied for comparison on examples from other geographical backgrounds. The model developed by Perkins, Brown, and Taylor (1996), which attempts to predict participation in community organizations in New York, Baltimore, and Salt Lake City, explains 28 percent of the variance of participation at individual level and 52 percent at block level. The model built by Long and Perkins (2007) to predict sense of community in New York explained 39 percent of the variance of the outcome variable at

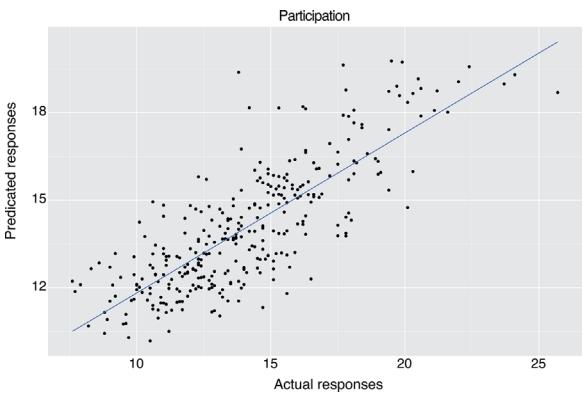



**Figure 5.** Participation (NI 003): Plot of the Predicted Responses to the Actual. *Note*: Different scalefrom sense of community, Figure 4.

individual level and 68 percent at block level. However, both these models include data from surveys organized on samples at local level, differentiating from our approach, which aims at using nationwide available data, in order to avoid the local surveys' shortcomings. 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 U.K. national surveys currently organized do not provide reliable data at this level, therefore, locally organized 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 the Prediction Techniques subsection, 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. Following, we discuss the results obtained for the two dimensions analyzed.

*Sense of Community.* As for sense of community, level of deprivation and the proportion of married people in the neighborhood are identified as the most important predictors, followed by "gender, age, household income, ethnicity, and cohabitation with a partner" (Sengupta et al., 2013, p. 39). 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 United Kingdom, the percentages of homeowners and of people providing unpaid care in the neighborhood may be associated with the relevance of place attachment and social networks in determining sense of community, as Long and Perkins (2007) reports. The role of vehicle and violent crimes in predicting sense of community is stated by Sherrieb et al. (2010), 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 Sengupta et al. (2013), 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.

*Participation.* Although "socioeconomic status by itself has no positive or negative effect on participation" (Dekker, 2007, p. 370), 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 (Alesina & La Ferrara, 2000; Rupasingha, Goetz, & Freshwater, 2006), 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 (Dekker, 2007; Rupasingha et al., 2006). The importance of the share of people living in private rented houses may be seen in agreement with what stated by Dekker (2007), if we consider it as a "negative" of the proportion of owner occupiers.

*Differences With Other Approaches.* To further evaluate our models, we trained multiple linear regression models using the same data sets. In order to reduce the chances of overfitting, we evaluated the models by using a 10-fold cross-validation. A larger number of folds would have entailed too small test sets. In both cases, the models performed slightly better than the ones trained with Random Forests: the sense of community model yielded $R^2 = 84.3$ percent and MSE 9.1, while for participation, the results were $R^2 = 74.9$ percent and MSE 3.2. It is worth to remind here that OOB estimates are deemed to be more conservative in reporting the accuracy of a model (Strobl et al., 2009b), even overestimating the error rate in case of data sets with number of variables larger than the number of instances (Mitchell, 2011). However, compared to regression models, Random Forests is able to deal with nonlinear data (Strobl, Boulesteix, Zeileis, & Hothorn, 2007), being therefore, suitable for a wider range of problems, while at the same time offering more human understandable results than other machine-learning approaches, such as neural networks or SVMs. Moreover, our results should be analyzed under a social science perspective, to be understood in depth. As pointed out by Berk (2006, p. 289), "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. This implies the need of a further study about the meaning of the differences between the indicators of participation and sense of community found in the literature using conventional statistics and the ones identified here, with regard to CC building. Moreover, it would be worth to investigate the relationships occurring between participation and sense of community, and the important variables aforementioned. This would be of particular interest, since the importance values provided by the Random Forests algorithm do not provide any description of how a variable influences the predicted outcomes.

## *Time*

CC dimensions are often measured to assess how they change during the implementation of a program, such as in MacLellan-Wright et al. (2007). Although

we used data collected over a long time span (2008–11), the measures provided by our models are not directly suitable. The majority of the data sets we used are from the 2011 Census. Censuses in the United Kingdom are organized every 10 years, therefore, other data sources should to be found to produce updated measures between one census and another.

## 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 LA 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 neighborhood level measures. Further work to build more geographically accurate models should then rely on other sources, such as locally organized 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 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 were the identification of data sets 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 dimensions 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 the Data Selection Criteria subsection 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, that is, of the predictivity of the variables.

**Alessandro Piscopo** is a Ph.D. Candidate at the University of Southampton [A.Piscopo@soton.ac.uk].
**Ronald Siebes, Ph.D.,** is a Senior Researcher at the VU university Amsterdam.
**Lynda Hardman, Ph.D.**, is a member of the management team of Centrum Wiskunde & Informatica (CWI) and part-time full Professor of Multimedia Discourse Interaction at the University of Utrecht.

## Notes

1. Unless differently specified, the terminology adopted in this subsection follows closely Friedman (1998).
2. http://data.gov.uk/dataset/ni-002-percentage-of-people-who-feel-that-they-belong-to-their-neighbourhood, consulted on August 14, 2016.
3. http://data.gov.uk/dataset/ni-003-civic-participation-in-the-local-area, consulted on August 14, 2016.
4. http://discover.ukdataservice.ac.uk/catalogue/?sn=6519, consulted on August 14, 2016.
5. In the R package used, the name given to this parameter was *minbucket*.
6. http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/table-types/index.html, consulted on August 14, 2016.

## References

Alesina, A., and E. La Ferrara. 2000. "Participation in Heterogeneous Communities." *The Quarterly Journal of Economics* 115 (3): 847–904.

Barbella, D., S. Benzaid, J.M. Christensen, B. Jackson, X.V. Qin, and D.R. Musicant. 2009. 'Understanding Support Vector Machine Classifications via a Recommender System-Like Approach." In *Proceedings of The 2009 International Conference on Data Mining, DMIN 2009*, eds. R. Stahlbock, S.F. Crone, and S. Lessmann. Athens, GA: CSREA Press.

Berk, R. 2006. "An Introduction to Ensemble Methods for Data Analysis." *Sociological Methods & Research* 34 (3): 263–95.

Breiman, L. 2001. "Statistical Modeling: The Two Cultures (With Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231.

Chainey, S. 2008. "Identifying Priority Neighbourhoods Using the Vulnerable Localities Index." *Policing* 2 (2): 196–209.

Dekker, K. 2007. "Social Capital, Neighbourhood Attachment and Participation in Distressed Urban Areas. A Case Study in The Hague and Utrecht, the Netherlands." *Housing Studies* 22 (3): 355–79.

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

Genuer, R., J. Poggi, and C. Tuleau-Malot. 2010. "Variable Selection Using Random Forests." *Pattern Recognition Letters* 31 (14): 2225–36.

Goodman, R., M. Speers, K. McLeroy, S. Fawcett, M. Kegler, E. Parker, S. Smith, T. Sterling, and N. Wallerstein. 1998. "Identifying and Defining the Dimensions of Community Capacity to Provide a Basis for Measurement." *Health Education & Behavior* 25 (3): 258–78.

Gutiérrez, N., R. Hilborn, and O. Defeo. 2011. "Leadership, Social Capital and Incentives Promote Successful Fisheries." *Nature* 470 (7334): 386–89.

Liberato, S., J. Brimblecombe, J. Ritchie, M. Ferguson, and J. Coveney. 2011. "Measuring Capacity Building in Communities: A Review of the Literature." *BMC Public Health* 11 (1): 850.

Long, D., and D. Perkins. 2007. "Community Social and Place Predictors of Sense of Community: A Multilevel and Longitudinal Analysis." *Journal of Community Psychology* 35 (5): 563–81.

MacLellan-Wright, M., D. Anderson, S. Barber, N. Smith, B. Cantin, R. Felix, and K. Raine. 2007. "The Development of Measures of Community Capacity for Community-Based Funding Programs in Canada." *Health Promotion International* 22 (4): 299–306.

Mashhadi, A., G. Quattrone, and L. Capra. 2013. "Putting Ubiquitous Crowd-Sourcing Into Context." In *Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13)*, New York, NY: ACM, 611–22.

McMillan, D., and D. Chavis. 1986. "Sense of Community: A Definition and Theory." *Journal of Community Psychology* 14 (1): 6–23.

Mitchell, M.W. 2011. "Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters." *Open Journal of Statistics* 1: 205-211.

Perkins, D., B. Brown, and R. Taylor. 1996. "The Ecology of Empowerment: Predicting Participation in Community Organizations." *Journal of Social Issues* 52 (1): 85–110.

Press, M. 2009. "Dimensions of Community Capacity Building: A Review of Its Implications in Tourism Development." *Journal of American Science* 5 (8): 172–80.

Quercia, D., D.O. Seaghdha, and J. Crowcroft. 2012. "Talk of the City: Our Tweets, Our Community Happiness." Paper presented at the International AAAI Conference on Web and Social Media, North America, May. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4611.

Rupasingha, A., S.J. Goetz, and D. Freshwater. 2006. "The Production of Social Capital in US Counties." *The Journal of Socio-Economics* 35 (1): 83–101.

Sengupta, N., N. Luyten, L. Greaves, D. Osborne, A. Robertson, G. Armstrong, and C. Sibley. 2013. "Sense of Community in New Zealand Neighbourhoods: A Multi-Level Model Predicting Social Capital." *New Zealand Journal of Psychology* 42 (1): 36–44.

Sharifuddin, N.S.M., M.S.M. Zahari, M. Aizuddin, and M.H. Hanafiah. 2015. Is the Sense of Community Towards Participation in Tourism Development Among the Minorities in Multiracial Countries the Same? World Academy of Science, Engineering and Technology. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering* 8 (11): 3699–707.

Sherrieb, K., F. Norris, and S. Galea. 2010. "Measuring Capacities for Community Resilience." *Social Indicators Research* 99 (2): 227–47.

Simmons, A., R. Reynolds, and B. Swinburn. 2011. "Defining Community Capacity Building: Is it Possible?" *Preventive Medicine* 52 (3): 193–99.

Siroky, D. 2009. "Navigating Random Forests and Related Advances in Algorithmic Modeling." *Statistics Surveys* 3: 147–63.

Statnikov, A., L. Wang, and C.F. Aliferis. 2008. "A Comprehensive Comparison of Random Forests and Support Vector Machines for Microarray-Based Cancer Classification." *BMC Bioinformatics* 9: 319.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics* 8: 25.

Strobl, C., T. Hothorn, and A. Zeileis. 2009a. "Party On!" *The R Journal* 1 (2): 14–7.

Strobl, C., J. Malley, and G. Tutz. 2009b. "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests." *Psychological Methods* 14 (4): 323–48.

Venerandi, A., G. Quattrone, L. Capra, D. Quercia, and D. Saez-Trumper. 2015. "Measuring Urban Deprivation From User Generated Content."In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15). New York, NY: ACM, 254–64.

Verikas, A., A. Gelzinis, and M. Bacauskiene. 2011. "Mining Data With Random Forests: A Survey and Results of New Tests." *Pattern Recognition* 44 (2): 330–49.