

Understanding urban gentrification through Machine Learning: Predicting neighbourhood change in London

Abstract

Recent developments in the field of machine learning offer new ways of modelling complex socio-spatial processes, allowing us to make predictions about how and where they might manifest in the future. Drawing on earlier empirical and theoretical attempts to understand gentrification and urban change, this paper shows it is possible to analyse existing patterns and processes of neighbourhood change to identify areas likely to experience change in the future. This is evidenced through an analysis of socio-economic transition in London neighbourhoods (based on 2001 and 2011 Census variables) which is used to predict those areas most likely to demonstrate ‘uplift’ or ‘decline’ by 2021. The paper concludes with a discussion of the implications of such modelling for the understanding of gentrification processes, noting that if qualitative work on gentrification and neighbourhood change is to offer more than a rigorous post-mortem then intensive, qualitative case studies *must* be confronted with—and complemented by—predictions stemming from other, more extensive approaches. As a demonstration of the capabilities of Machine Learning, this paper underlines the continuing value of quantitative approaches in understanding complex urban processes such as gentrification.

Keywords

London, neighbourhood change, gentrification, principal components, machine learning, random forests, census, quantitative geography

Introduction

The application of quantitative methods to the study of neighbourhood change in general—and gentrification in particular—still has something of a controversial air. Despite some of the most-cited works in the field utilising quantitative methods to either measure the ‘rent gap’ between actual and potential housing rents (*e.g.* Ley 1986; Clark 1988) or demonstrate socioeconomic change through census analysis (*e.g.* Atkinson 2000; Hamnett 2003), the majority of literature on gentrification now shuns quantitative analysis in favour of qualitative assessments of neighbourhood change based on media analysis, interviews, ethnography and other forms of observational data collection. In part, this is because of the limitations of secondary data for capturing the dynamics of urban processes occurring at a local level (Watt 2008), but this is often coupled with a suspicion that ‘official’ statistics relating to neighbourhood change describe patterns but obfuscate underlying processes of class change (Slater 2009).

Consequently, in most contemporary accounts, intensive and qualitative methods are the favoured means of exploring urban gentrification; however, the privileging of such methods is not without risks since, as Barton (2016: 92) points out, “qualitative strategies for identifying gentrified neighbourhoods may overlook areas that experienced similar changes to those more widely recognised as gentrified.” Focusing on New York, Barton (2016) and others (*e.g.* Bostic and Martin, 2003; Freeman, 2005) use regression methods to reveal a much larger number of census tracts where gentrification seems to have occurred than those generally highlighted in the literature. This suggests that the academic and media preoccupation with Brooklyn and Manhattan districts experiencing obvious social and cultural change (*e.g.* a transition from black to white occupation and the associated rise of ‘hipster’ stores) distracts from a wider appreciation of the situation across the five Boroughs.

In other cities, a similar privileging of select ‘signifying locations’ appears equally evident, with certain neighbourhoods repeatedly attracting the researcher’s gaze; as Neal *et al.*

1
2
3
4 (2016) wittily put it: ‘You can’t move in Hackney without bumping into an anthropologist’.
5
6 Indeed, recent analyses of London have fixated on specific parts of the East End (e.g.
7 Harris 2012 on Hoxton; Watt 2008 on Stratford; and Butler *et al.* 2013 on Hackney) or
8 South London (e.g. Jackson and Benson 2014 on Peckham; Mavromattis 2012 on Brixton),
9
10 potentially ignoring other neighbourhoods where significant change is occurring.
11
12 Quantitative and multivariate analysis across a range of neighbourhoods hence appears
13
14 important for grasping the bigger picture and, more importantly, it appears such methods
15
16 could predict where the ‘gentrification frontier’ might move to next (see Chapple 2009).
17
18

19
20 The work presented here provides a quantitative analysis of this kind and is motivated by
21
22 the emergence of ‘machine learning’ techniques (hereafter: ML) that have the capacity to
23
24 learn from, and make predictions about, observations in large data sets without being
25
26 explicitly programmed with a model of how to do so. We will detail our specific approach
27
28 later, but suffice to say here that most ML approaches incorporate some form of
29
30 optimisation (a measure of whether the predictions are getting better or worse), alongside
31
32 phases of training (in which the algorithm learns how to make predictions based on
33
34 ‘existing’ data) and testing (in which results are tested for robustness using ‘new’ data).
35
36

37
38 While such methods will not necessarily lead to new theories of gentrification on their own,
39
40 in this paper we suggest that they can indicate possible *trajectories of neighbourhood*
41
42 *change*, something that is particularly important in theory development (Owens 2012). We
43
44 explore this contention by using the ‘random forests’ algorithm to tease out the trajectories
45
46 of 4,835 London neighbourhoods between 2001 and 2021, based on analysis of social,
47
48 economic and environmental variables. The contribution of this paper to gentrification
49
50 debates is not, however, solely methodological (*i.e.* showing how we can use ML methods
51
52 to predict urban change) but also empirical (*i.e.* mapping shifts in London’s ‘gentrification
53
54 frontier’ via a fine-grained analysis of neighbourhood change).
55
56
57
58
59
60

Modelling neighbourhood change

It has been suggested that gentrification needs to be understood as a neighbourhood-level phenomenon involving not just an increase in the value of an individual property, but a simultaneous uplift in the values of comparable properties across a given neighbourhood (O'Sullivan 2002). In classic theories of gentrification this uplift is associated with the arrival of new, wealthier populations and the displacement of existing inhabitants, alongside improvements to the housing stock that register this socio-economic transition (Atkinson 2000). Alternative theories suggest that improvements to the built environment can also occur via *marginal gentrification* caused by the arrival of culturally-rich—though not necessarily affluent—populations, such as artists and students (Hochstenbach *et al.* 2015), and via *incumbent upgrading* by longer-term residents (Van Criekingen and Decroly 2003). Owens (2012:347) operationalised these in a quantitative context using the concept of neighbourhood 'Socio-Economic Status' (SES) change: we adopt this given it potentially reveals change-processes other than gentrification and displacement *per se*.

Notwithstanding the risk that some neighbourhood processes occur at a granular level that cannot be 'seen' through quantitative data (Barton, 2016: 99), there remains the challenge of defining a neighbourhood in the first place. Here, there are a host of overlapping definitions available, but for our purposes the one advanced by Galster (2001: 2112) offers a suitable starting point: "the bundle of spatially-based attributes associated with clusters of residences, sometimes in conjunction with other land uses." While this does not establish neighbourhoods as discrete, bounded entities (*i.e.* it does not unambiguously state how big or small a neighbourhood is), it provides a basis for defining neighbourhoods on different spatial scales through the 'bundling' of attributes. In effect, Galster defines a set of 'domains' within which neighbourhood-ness is constructed, namely: urban morphology; mobility and utility infrastructures; demography; class; tax and public services; the environment; proximity to facilities (both recreational and employment-based); political networks; degree of social interaction; and sentiment (*i.e.* place attachment).

1
2
3
4 In a US context, Van Criekingen and Decroly (2003:2457) employed indicators of
5 deprivation, upgrading of the built environment, social status, population, and income
6 change to classify neighbourhoods on this basis. Here, there are obvious parallels to
7 geodemographic analyses of the type underpinning the operationalisation of the 2001 and
8 2011 Output Area Classifications in the UK (Vickers and Rees 2007; Gale and Longley
9 2011; Gale 2014; and see also Li and Xie 2018 on the clustering of US census data, 1970–
10 2010). But while geodemographics uses area attributes to assign neighbourhoods to groups
11 (*i.e.* clusters), we use these attributes to predict an outcome.
12
13
14
15
16
17
18
19
20

21 **Contextualising machine learning in urban studies**

22
23 To date, ML has most commonly been employed in physical geography where it is often
24 used in conjunction with remotely-sensed data to classify landforms (Xiao 2016). Recently,
25 the use of ML in topics of interest to human geographers—such as changes to the fabric of
26 cities, the prediction of transport modality, detection of deprivation, and population
27 prediction—has grown rapidly as well (*e.g.* Arribas-Bel *et al.* 2011; Arribas-Bel *et al.*
28 2017; Donaldson and Storeygard 2016; Hagenauer and Helbich 2017; Naik *et al.* 2017; Liu
29 *et al.* 2017; Santibanez *et al.* 2015; Stevens *et al.* 2015). Revisions to classical regression
30 techniques have also yielded geographically-aware ML tools such as Spatially-Filtered
31 Ridge Regression (Fan *et al.* 2016), and derived probability transitions aiding
32 understanding of the evolution of regional income disparities (Rey 2014).
33
34
35
36
37
38
39
40

41 Because ML differs radically from approaches commonly employed by social science
42 researchers it is worth clarifying what ML can—and cannot—accomplish. The most
43 obvious difference to conventional methods is simply one of scale: ML algorithms not only
44 tackle very ‘long’ data sets containing many rows, they also tackle very ‘wide’ ones
45 incorporating many correlated variables (as intercorrelation does not impact ML
46 approaches in the same way as traditional multivariate analysis, meaning methods can
47 make better use of the full extent of the data). Clearly, a not coincidental reason for the rise
48 of ML is the growing availability of ‘big data’ about human society: telephone usage
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 (Reades and Smith 2014), vehicle licensing (Lansley 2016), public transit smartcard usage
5 (Zhong *et al.* 2014), and even taxi trips (Manley *et al.* 2015) are all amenable to analysis.
6
7 Of course, many cultural aspects remain ‘off the radar’ (Barton 2016:94), but in the context
8 of neighbourhood change, social media such as Twitter or Instagram, and even Tripadvisor
9 reviews, can offer useful proxies (see Boy and Uitermark 2016; Hristova *et al.* 2016; Zukin
10 *et al.* 2017).
11
12
13
14

15 Unlike conventional statistical methods, ML approaches are not necessarily concerned with
16 causality, being primarily concerned with *utility*. The online retailer Amazon, for instance,
17 does not care why there is a strong relationship between two books in its customers’
18 purchasing patterns, only whether they can influence the customer to buy the second book.
19 As Wyly (2014:681) puts it: “The capitalist correlation imperative is clear: spurious
20 correlation is fine, so long as it is *profitable* spurious correlation.” The capacity of modern
21 corporations to ‘consume’ large volumes of data with which to make profitable predictions
22 is *one* outcome of the rise of ML and ‘big data’, but the availability and openness of these
23 tools—they are not ‘black boxes’ to quite the extent that Dalton and Thatcher (2015) appear
24 to believe—means that researchers are now in a position to create ‘early warning systems’
25 (Chapple 2009; Chapple and Zuk 2016; Steif *et al.* 2017) to alert residents, representatives,
26 and policy-makers to incipient changes in an area’s social and economic dynamics.
27
28
29
30
31
32
33
34
35
36

37 This noted, the research undertaken in this article explores neighbourhood change in
38 London using 166 variables across transport, housing, demographics, income and wealth,
39 amenity, and occupational domains. Ultimately, this article does not seek to provide new
40 insights into the root causes of gentrification—these have been amply covered elsewhere in
41 the literature (*e.g.* Davidson and Lees 2005; Hamnett 1984; Redfern 1997, 2003; Zukin *et*
42 *al.* 2009)—but uses contemporary ML techniques to help select features (*i.e.* variables)
43 from the available data in one time period that might be useful for predicting status change
44 in the next, and to use the outputs of our model to foster debate about the changing urban
45 geographies of the Greater London Authority (which includes 32 London Boroughs and the
46 City of London).
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Methodology

As we noted above, with the principal exception of work by Hamnett (1983, 2003, 2009, 2015), census data has been sparingly used in studies of gentrification and neighbourhood change in the UK. In contrast, North American studies have more frequently used secondary data (*e.g.* Barton 2016; Bostic and Martin 2003; Freeman 2005, 2009; Meligrana and Skaburskis 2005; Owens 2012). In one early study, Melchert and Naroff (1987:681) employed logistic regression on data for Boston, MA to establish that ‘amenity, social, housing and economic variables [*have*] predictive capabilities [*that are*] quite substantial... [*indicating*] that the general context of a neighbourhood is of far greater significance than individual groups of characteristics.’

The utility of regression may, however, be severely impacted by collinearity (such as might be expected between education and income, or income and property prices). This interdependence is often associated with instability in the model thanks to the ‘inflation’ of coefficients such that some inputs gain in significance at the expense of other, equally important but partially correlated, variables. Stepwise regression was an early computational means of trying to cope with this challenge, but has now been superseded by more robust approaches—generically and collectively referred to as ML—and it is for this reason that this paper explores the potential of ML for advancing understanding of neighbourhood change.

There are obvious limits to how fully we can document our method, so we focus here on the key steps. However, an important overarching consideration is the importance of open, replicable research (*e.g.* Singleton *et al.* 2016); by using both open data and open source code, we enable replication (Brunsdon 2016) by researchers, activists, policymakers, or even real-estate developers. Indeed, our analysis employs only open data (from the 2001 and 2011 UK Census of Population and the London Data Store - an extensive open data portal). Any reader who disagrees with our methodological choices is also free to adapt the code since this is also freely available—for downloading, revision, and (re)running—as a series of Python-based ‘notebooks’ on the GitHub code-sharing web site.

Data Assembly

A predictive model of neighbourhood change needs two sets of variables: those that measure the status of a neighbourhood, and those that help us predict changes to come. But even before we get to variable selection, it should be noted that the quantitative analysis of neighbourhoods presents several practical challenges, not least of which is the selection of an appropriate geographical scale. Lauria and Stout (1995) have argued that a block-by-block analysis is essential, but cutting against this claim are two inter-related issues: firstly, that fine-scale data are often considered highly sensitive and suppressed from census outputs; and, secondly, that natural variation between smaller areas yields statistically significant—but not actually meaningful—fluctuation (*i.e.* noise). A good example of the latter would be property prices: at the *street* level, the ‘average house price’ in any given year might be based on a single transaction for an unrepresentative property! Conversely, larger areas generally lack a sense of cohesion and shared identity that we might associate with a similar quality of life, housing conditions, access to services and so on, and necessarily tend to smooth out variation to undermine the detection of change.

Putting these contradictory effects together suggests it is easiest to work with intermediate or meso-scale data; fortunately, the Office for National Statistics (ONS) provides one such grouping in the Lower Layer Super Output Area (LSOA) (broadly similar to a US census tract). The LSOA contains between 1,000 and 3,000 inhabitants living in between 400 to 1,200 households: a geography small enough that even modest changes in the makeup of an area should show up, but large enough that the sample size of each is statistically robust. Whilst data is available at both finer (*e.g.* Output Areas) and coarser scales (*e.g.* wards or Middle Layer Super Output Area), work in the UK concludes that LSOAs exemplify the characteristics of spatial proximity and social homogeneity which are revealing of “neighbourhood effects” (van Ham *et al.* 2012).

So although LSOAs are statistical units rather than an empirical reality, they are broadly coterminous with the kinds of environments that appear important in giving residents both a sense of identity and a context for everyday life. In fact, up to a point LSOAs are

1
2
3
4 deliberately constructed to contain a broadly-consistent housing type and demography (see
5
6 Cockings *et al.* 2011). Analysis at this scale hence provides the main basis for
7
8 understanding the production of neighbourhoods as socially meaningful and physically
9
10 distinctive urban spaces in London (Sturgis *et al.* 2014).

11 12 ***Calculating Scores*** 13

14 If we begin by assuming that the indicators identified by Van Criekingen and Decroly
15
16 (2003) are sufficiently comprehensive then—drawing on Owens (2012)—we can use four
17
18 variables to measure neighbourhood status: household income (using the modelled median
19
20 value in each neighbourhood¹), property sale value (also using the median value),
21
22 occupational share (the percentage of the neighbourhood’s residents in the ‘top’
23
24 occupational classes), and qualifications (the percentage of residents achieving NVQ Level
25
26 4 or above). Though private sector rents would have been a useful complement to sales,
27
28 historical data for this domain is very limited in the UK.

29 To train the ML algorithm to predict neighbourhood change we need to combine these four
30
31 variables into a singular measure of ‘socioeconomic status’. Since we are working with a
32
33 long but fairly-narrow data matrix, Principal Components Analysis (PCA) is an obvious
34
35 choice as it will yield just four components: by taking just the first one we capture the
36
37 majority of the variation in the input data using a single numeric value. This will
38
39 necessarily cause *some* loss of detail about neighbourhoods because we do not retain any of
40
41 the subsidiary components, but we can quantify this loss using the percentage of variance
42
43 explained by each component (this is also the approach taken by Owens, 2012, following
44
45
46
47
48
49

50 ¹ Household income is not normally available at the LSOA scale in Britain, but the Greater
51
52 London Authority undertook a modelling project incorporating access to restricted data to
53
54 produce this for London.
55
56
57
58
59
60

1
2
3
4 Morenoff and Tienda, 1997). Additionally, we apply PCA simultaneously to both census
5 years to avoid the problem that scores for different years are not directly comparable.
6
7

8
9 The construction of these scores necessarily entailed decisions about the re-scaling of
10 variables since differences in magnitude could allow one dimension to dominate (*e.g.* house
11 prices vs. share of high-qualifications). Simple unit scaling (*i.e.* remapping the range of
12 each variable to the scale 0–1) is unlikely to address this problem because the existence of
13 ‘heavy tails’ would lead to the bunching of the data at one end of the scale. Equally, since
14 house prices and incomes are also highly-skewed, the mean is unlikely to be a robust
15 measure of centrality. Robust standardisation using the median and Inter-Quartile Range
16 (IQR) addresses both issues: it preserves outliers while producing comparable scales for the
17 bulk of the data. In our testing, this approach yields the most consistent performance and
18 was applied to all score dimensions. More aggressive, non-linear transformations are
19 possible for extreme distributions prior to this step, but these typically lead to the loss of
20 information about the magnitude of outliers or the balance between dimensions in the
21 score.² To ensure that the two census years are directly comparable we apply the same
22 transformation to both.
23
24
25
26
27
28
29
30
31
32

33 34 ***Selecting Predictor Variables*** 35

36 In line with previous work in this area we attempted to select variables from a range of
37 categories including: Housing, Households, Work, Travel and Amenity. This set is far from
38 exhaustive, and the use of more built environment and amenity features (*e.g.* schools)
39 would be one obvious areas for improvement; however, these nonetheless encompass the
40 principal areas on which work on gentrification and neighbourhood change have focussed.
41 Rather than reproduce the full list of 166 variables, readers are invited to access the
42 additional details in the online repository. Of course, the alert reader will have realised that
43
44
45
46
47
48
49
50

51
52 ² The code on GitHub also allows readers to apply Box-Cox and Log transformations to
53 these data to explore the impact of scoring changes on the overall results.
54
55
56
57
58
59
60

1
2
3
4 some variables will necessarily play a role in both scoring *and* prediction so it is inevitable
5 that the scores will be correlated with property price, income, skills, and occupation data.
6
7

8 9 ***Relative vs Absolute Measures***

10
11 Lees (2000:403) argues both ‘contextuality and scale are significant’ in gentrification
12 research, implying the need to incorporate *relative* measures of change as part of any
13 neighbourhood analysis. For instance, given trends in London it is entirely conceivable that
14 an area can experience ‘ascent’ (*i.e.* an absolute ‘improvement’ in its score) but at a lower
15 rate than its neighbours (*i.e.* a relative ‘decline’). Equally, if gentrification is understood in
16 terms of in-movers having a multiple of the current residents’ median income, then ‘super-
17 gentrification’ (Lees 2003; Butler and Lees 2006) may appear quite similar to ‘plain old’
18 gentrification in a relative sense. This is a ‘feature’ and not a ‘bug’ of this approach: we can
19 use relative to change to effectively classify *both* as forms of gentrification even if they
20 differ in an absolute sense.
21
22
23
24
25
26
27
28

29
30 On a practical note, raw values can also be problematic for ML because ‘decision
31 boundaries’—the thresholds used for regression or classification—will almost certainly
32 shift over time. For instance, if crime generally falls across London between 2001 and 2011
33 then a ‘low’ rate of neighbourhood crime in one Census year is *not* the same as a low rate
34 in the next Census year. Consequently, judged in absolute terms many more areas will
35 appear to have become attractive to gentrifiers even if the relative differences between
36 areas remain substantial. Similarly, even if the relative proportions for each demographic
37 group in city remain the same, an expansion in the absolute number of households could
38 lead to housing stress if supply fails to keep up with demand (Hamnett 2015:244).
39
40
41
42
43
44
45

46 47 ***Random Forests***

48
49 Random Forests (see James *et al.* 2013 for a systematic introduction) are a particularly
50 versatile and robust form of non-parametric ML, able to perform both classification
51 (assigning observations to classes) and regression (predicting values from observations)
52 tasks quickly, without much tuning and with minimal bias (Breiman 2001). The term
53 ‘random’ originates from the way that Random Forests (RFs) employ random subsets of the
54
55
56
57
58
59
60

1
2
3
4 available dimensions (*i.e.* variables) to avoid the risk of over-fitting. RFs are *ensemble*
5 methods, meaning they aggregate the output of a large number of *decision trees*—many
6 trees yields one forest—and so can cope with complex, non-linear decision boundaries. We
7 tackle this terminology and its import below.
8
9
10

11
12 To understand more fully how this approach works, let us take a simple decision tree:
13 anyone who has played the game Twenty Questions has employed a decision tree since,
14 with each new question, the player divides the ‘answer space’ into two smaller spaces, one
15 of which is excluded from subsequent consideration (*e.g.* is it bigger than a shoebox? Is it
16 alive?). Shallow trees employing a relatively short sequence of questions can uniquely
17 identify a single ‘thing’ from a very large number of possible ‘things’ remarkably quickly.
18 Twenty Questions is a classification problem, but this approach can *also* be used for
19 regression: is it before 10am? After 8am? Is it a weekday? A highway? Applying these
20 questions to some movement data we can predict rush hour volumes. James *et al.*
21 (2013:306) describe the function of a tree as ‘prediction via the stratification of the feature
22 space’ using a two-step process: the predictor space is divided into a set of ‘distinct and
23 non-overlapping regions’ and for every observation falling into a given region we make the
24 same prediction (usually the mean of observations from the data used when growing the
25 tree). We will unpack this statement later, but by way of an illustration we show in **Error!**
26 **Reference source not found.** *part* of an actual tree—one of the many grown by the
27 Random Forest on the data—created as part of this research.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

49 [Insert Figure 1 here]

50 Although trees can be manually created using expert knowledge, their growth can also be
51 automated using a ‘heuristic’: typically, the computer selects the dimension that best-
52 enables it to split the data set into two dissimilar groups. At each ‘node’ (branch in the tree)
53 we deal with progressively smaller subsets of the data and this process continues down each
54
55
56
57
58
59
60

1
2
3
4 branch until some stopping point—termed a ‘leaf’—is reached. The RF grows each tree on
5 a *randomly*-selected subset (S) of all dimensions (D); these subsets overlap such that trees
6 use similar, but not identical, subsets of D . Randomness is then used *a second time* since
7 the tree is further restricted to considering a random subset of S with which to split the
8 ‘remaining’ data at each node. This approach decorrelates the trees by preventing an over-
9 reliance on any one variable and so helps to prevent over-fitting of the data.
10
11
12
13
14

15
16 The many trees in the forest then ‘vote’ as an *ensemble* on their preferred class or predicted
17 value, but the poorly-performing trees tend to cancel each other out (noise) while the useful
18 ones (signal) carry the day. In fact, our model goes further than this by employing the
19 computationally-efficient ‘extremely randomised trees’ (Guerts *et al.* 2016): this not only
20 employs randomly selected dimensions, it also uses random ‘cut points’ for each split. The
21 prominence of randomness in this method might seem strange to some readers, but in
22 statistical terms it is highly robust.
23
24
25
26
27

28 ***Training & Testing***

29
30
31 An important component of most ML approaches is the incorporation of training and
32 testing regimes: we train the algorithm on a random subset of the full data set, and then test
33 its performance against the portion of the data set not already used. K -fold cross-validation
34 is a common approach: the full data set is split into k ‘folds’, each of which is used $k-1$
35 times as part of the training data set, and *once* as the testing data to be predicted. This has a
36 significant impact on the model’s overall bias and helps to ensure that outliers do not
37 unduly impact the model. Here, randomisation again helps improve the robustness of our
38 predictions.
39
40
41
42
43
44

45 ***Hyperparameter Tuning***

46
47
48 Finally, and in common with many ML approaches, we still need to define how the
49 algorithm should ‘learn’ about the data and gauge its performance. The RFs learning
50 process is governed by ‘hyperparameters’ and the most important considerations are:
51
52
53
54
55
56
57
58
59
60

- That more estimators (trees) may yield more nuanced predictions but can overfit some data.
- That trees can be grown to any depth, but specifying a maximum depth reduces the risk of overfitting with ‘deep’ trees.
- That the minimum size of leaves should normally be a small number (higher resolution predictions) but can also lead to overfitting with some data.
- That reducing the proportion of features used by a tree helps to manage correlation with other trees by reducing their overlap.

Together, these hyperparameters constitute a ‘space’ that can be systematically explored as part of the model configuration process. We divide this space into a grid and test every combination of hyperparameters using the k -fold training approach set out above. We can compare the performance of each configuration using the Mean Squared Error or Mean Absolute Error of the predictions. It is also possible to generate a R^2 value, although using this metric for direct model comparison is considered problematic.

Neighbourhood change in London 2001-11

To recap, we are using a model built on the characteristics of LSOAs from the 2001 Census to ‘predict’ the 2011 scores, and then use same model with the 2011 Census data to predict outcomes in 2021. Obviously, predictions remain extrapolations (however sophisticated), and predicting the future is always fraught with difficulty: Hamnett (2003) expected that Clapton in East London would prove resistant to gentrification but it is an area that is now very much on—or even behind—the gentrification frontier (Holland 2012).

Ideally, we would take a longer-term view but, unfortunately, compatible census data is not available to catch the initial waves of gentrification in Islington and Notting Hill (*e.g.* Glass 1964), but we would expect any analysis of neighbourhood change in London using 2001–2011 data to pick up signs of status changes in areas such as London Fields, Dalston, Brixton and Peckham (Butler and Robson 2001; Benson and Jackson 2017). It might, of course, also show up changes associated with super-gentrification in neighbourhoods that

1
2
3
4 experienced gentrification in earlier periods (see Butler and Lees 2006 on Barnsbury), as
5 well as areas demonstrating forms of incumbent improvement where displacement has not
6 been a significant factor which is something that Freeman *et al.* (2015) suggest could well
7 apply in London.
8
9
10

11 ***Scoring Results***

12
13
14 Even after robust re-scaling, property prices and incomes ‘count’ for more than changes in
15 skills or occupational mix in our scores, and following PCA the percent of variance
16 explained by the first component (our score) is 78.8%. If we understand this as a way of
17 mapping the data onto new axes aligned with variation in the ‘data cloud’, then the
18 discarded components—accounting for 15.1%, 4.9% and 1.2% of variance respectively—
19 capture lesser variation that we can loosely term ‘noise’ even though they might, in the
20 round, still prove useful for prediction.
21
22
23
24
25
26

27 **Error! Reference source not found.** shows two axes of high property values emanating
28 from Central London—Southwest and North-Northeast—with ‘Billionaire’s Row’
29 (Bishop’s Avenue) on Barnet’s border with Haringey featuring prominently. In the context
30 of an ‘affordability crisis’ in London housing (see Hamnett and Reades 2018), the emphasis
31 on property price in our measurement of neighbourhood status encapsulates one of the main
32 mechanisms through which even fairly well-off residents are experiencing neighbourhood
33 change (Benson and Jackson 2017).
34
35
36
37
38
39
40
41
42

43 [Insert Figure 2 here]
44
45
46
47

48 ***Model Comparisons***

49
50
51 Hyperparameter tuning—optimising for Mean Squared Error (MSE)—yielded a RF with a
52 configuration of: 1,400 trees, 85% of features considered by each tree, no maximum tree
53 depth, and a minimum leaf size of two. Compared to traditional methods (**Error!**)
54
55
56
57
58
59
60

1
2
3
4 **Reference source not found.**), the RF shows improvements over both types of linear
5 regression even without tuning, but the tuned model outperforms multiple linear regression
6 by more than 10% across every measure.
7
8
9

10
11
12
13 [Insert Table 1 here]
14
15
16
17

18 However, the ultimate value of the model lies in how well it predicts the 2011 scores using
19 the 2001 data: a Pearson's r of 0.99 indicates that for most observations the forest performs
20 very well indeed. There *are* outliers of course, though it is reasonable to expect that major
21 property developments, as well as the 'decanting' of residents from council estates
22 undergoing redevelopment (*e.g.* Lees 2014), might transform individual neighbourhoods in
23 ways that no predictive model could anticipate.
24
25
26
27
28
29

30 *Predictor importance*

31
32 Before introducing the predictions in detail we examine which variables the model found
33 most important for predicting status change. A feature importance measure is automatically
34 generated by RFs and is best understood as the contribution of the variable to the model.
35 This metric is measured out of a theoretical maximum value of 1—so larger values mean
36 more useful variables—but with 166 variables it is impractical to show these in a table and
37 a visual representation has been used instead.
38
39
40
41
42
43
44
45

46 [Insert Figure 3 here]
47
48
49
50

51 is broadly consistent with hypotheses that relate to occupation and skills changes as drivers
52 of neighbourhood change (Hamnett 2015): work-related variables make up much of the
53
54
55
56
57
58
59
60

1
2
3
4 top-20, with long hours (for both men and women), skills and qualifications (both high and
5 low), and job flexibility (self-employment with and without employees, as well as
6 homeworking) all good predictors of neighbourhood status change. Immigration from the
7 Americas, 2001 EU members, and Oceania also show up in the top-30, suggesting that
8 global-scale inflows are also a useful predictor (see Butler and Lees 2006). Older buildings
9 remain attractive to in-movers (as hypothesised by Glass 1964 and many others), but rather
10 less expected is the fact that ‘DINKS’ (Dual-Income, No Kids) do not feature strongly,
11 though this is consistent with Karsten’s (2003) observation of a shift towards child-rearing
12 in the ‘Inner City’.
13
14
15
16
17
18
19
20
21
22

23 [Insert Figure 3 here]
24
25
26
27
28

29 *Trajectories of change*

30
31 Taking an overview, **Error! Reference source not found.**a shows the changing
32 distribution of scores over time, suggesting a flattening of the distribution whilst implying
33 continued status change likely to have a pronounced impact on the most affordable and
34 least-well off LSOAs. Note, however, that this trend is *not* expected to accelerate: **Error!**
35 **Reference source not found.**b predicts an overall slowing of the magnitude of change. The
36 neighbourhoods that have experienced the strongest change in 2001–2011 show
37 comparably less change in the subsequent period.
38
39
40
41
42
43
44
45
46

47 [Insert Figure 4 here]
48
49
50
51

52 The more interesting analysis, however, is a geographical one: where is change most
53 significant across the two-time periods? Since *everywhere* is experiencing status score
54
55
56
57
58
59
60

1
2
3
4 increase over the period 2001–2021 it is more useful to examine relative changes in the
5 *ranking* of LSOAs. We could have random fluctuations in the rankings based on very
6 minor differences in input variables, so it would be preferable to avoid taking ‘noise’ an
7 indicator of significant change. Accordingly, since the distribution of changes in rank was
8 broadly both symmetric and normal, these movements were grouped by standard deviation:
9 more extreme values are more likely to indicate meaningful change. Movements within ± 1
10 Standard Deviation are not shown in **Error! Reference source not found.** on the basis that
11 they are most likely to represent random fluctuation.
12
13
14
15
16
17
18
19
20
21

22 [Insert Figure 5 here]
23
24
25
26
27

28 Broadly, **Error! Reference source not found.** shows Inner East London—those areas near
29 the London Olympic development especially—‘catching up’ with *non-prime* West London.
30 This is *not* to suggest that West London has seen some sort of decline, only that it is
31 improving at a slower rate. ‘Prime London’ in Westminster and Kensington & Chelsea
32 obviously saw enormous gains in 2001–2011, but the significant changes were concentrated
33 towards the north ends of both boroughs where pockets of deprivation and un-upgraded
34 housing remain.
35
36
37
38
39
40

41 Running the predictions forward to 2021 (**Error! Reference source not found.**) sees these
42 concentrations disperse, though this should not be confused with an absence of change in
43 these areas. What is striking about the comparison with **Error! Reference source not**
44 **found.** is the shift outwards from Inner East London: a wedge of ‘uplift’ now extends out to
45 the traditionally working class boroughs of Havering, Waltham Forest, and Bexley. ‘Prime
46 London’ continues to pull away from the rest of the city in absolute terms, and we expect
47 the vestiges of deprivation in these boroughs to be wiped out by the ongoing redevelopment
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 of council estates in both Westminster and Kensington & Chelsea (Lees 2014; Minton
5 2017).
6
7

8
9 In contrast, there are areas of relative decline in the outer boroughs of Croydon, Harrow
10 and Hounslow implying that these are less likely to experience the changes and
11 displacements associated with improving levels of education and in-movers engaged in
12 higher-status work (see Leckie 2009 and Butler *et al.* 2013 on links between education and
13 gentrification in London). A further implication is that the uplift of the East End may well
14 be linked to displacement of the least well-off to Outer London (Travers *et al.* 2016)—
15 something that Freeman *et al.* (2015:2811) also see as a distinct possibility given both that
16 the poor are forced to move more frequently than the well-off, and that those moving into
17 gentrifying areas are nearly three times more likely to have a degree than those moving into
18 disadvantaged neighbourhoods.
19
20
21
22
23
24
25
26
27
28
29

30 [Insert Figure 6 here]
31
32
33

34 **Discussion & limitations**

35
36 For those who live in London, and who have the benefit of hindsight, some of these
37 predictions may appear self-evident: these areas are on most people's radar and might even
38 be seen to be areas where change has 'been and gone'. However, it is worth recognising
39 that the *preconditions* of these changes must have been in place by 2011 for these
40 predictions to be made and that, had we had access to this data *in 2011*, then we could have
41 made these predictions at that time! It is therefore possible to envision revisions to our
42 approach to incorporate more 'timely' data—such as from *Zoopla* (a property price
43 website) or *Twitter* (useful as a marker of cultural change)—to develop the kind of real-
44 time 'early warning system' anticipated by Chapple and Zuk (2016).
45
46
47
48
49
50
51
52

53 Although we have singled out Hamnett (2003) for his erroneous prediction of 'no change'
54 in Clapton (Hackney) there is, of course, no guarantee that we will do better. Nonetheless,
55
56
57
58
59
60

1
2
3
4 if studies of gentrification and neighbourhood change are to offer more than a rigorous
5 post-mortem then intensive case studies *must* be confronted with—and complemented by—
6 predictions stemming from other approaches. Indeed, we hope to be proven wrong in some
7 of our predictions, but explaining *why* we got it wrong should enrich understanding of the
8 factors influencing areas in transition. For instance, Lees (2000:398) has noted there is a
9 temporal aspect to change which means that the gentrifiers of today are not necessarily the
10 same as those of the 1980s, so a clear limitation of the approach is that the model links the
11 markers of change in 2011–2021 to those of 2001–2011. That said, it should also be
12 recognised that the algorithm is not impacted by our human propensity to simplify and
13 generalise, so while ML may be vulnerable to unforeseen behavioural change it is also
14 more subtle in terms of how it makes use of the available data.
15
16
17
18
19
20
21
22
23

24 Regardless, longer-term data going back to 1981 or 1991 would benefit our approach
25 substantially and enable us to explore the regeneration of the Docklands in the 1980s
26 (Foster 1999) alongside trends highlighted by Hamnett (2009). Unfortunately, we have no
27 equivalent to the US Neighbourhood Change Database (Barton 2016:7) which provides
28 comparable data across multiple Censuses, and changes in the classification of account and
29 small employers present additional challenges in using data of this vintage (Hamnett
30 2015:240–241). The absence of a gridded population surface on the Northern Irish model
31 (*e.g.* Martin *et al.* 2011) also limits longitudinal research because of incompatible zone
32 definitions; although the ‘PopChange’ project (Lloyd *et al.* 2016) is a promising step in this
33 regard it is insufficient in terms of both resolution and the variables available.
34
35
36
37
38
39
40
41

42 Another factor that we have not directly addressed in this paper is the influence of
43 neighbouring zones and ‘edge effects’: Redfern has argued that gentrification operates by a
44 diffusion process (1997:1337), and Kolko (2007) noted that the income of adjacent census
45 tracts might be a useful predictor of future neighbourhood change. It is likely that the
46 incorporation of, for example, spatial lags via Local Indicators of Spatial Association
47 (Anselin 1995) might improve our predictions. Moreover, change does not magically cease
48 at the edge of London’s administrative boundaries: we know that the past two decades have
49 been characterised by the increasing suburbanisation of poverty (Travers *et al.* 2016) and
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 would have liked to expand our analysis beyond the GLA boundary but income data is not
5 available at the LSOA scale outside of London.
6
7

8
9 There is, however, nothing to ultimately prevent us modelling the entire UK to search for
10 larger patterns of neighbourhood change such as rural in-migration or the impact of empty
11 second homes in areas such as Devon or Cornwall. Achieving this, however, will require
12 the development of a deeper understanding of the typologies of neighbourhood change
13 captured by the scoring metric through its interactions with the ML algorithm, something
14 we anticipate undertaking as a piece of follow-on work in due course.
15
16
17
18
19
20
21

22 **Conclusion**

23
24 Gentrification research remains mired in debates about cause and effect, and whether
25 displacement inevitably accompanies neighbourhood improvement (Hamnett 2003; Lees
26 2000; Freeman *et al.* 2015). Quantitative work has something to contribute here, showing
27 where status change is occurring and relating it to other variables in a way that generates
28 useful hypotheses about mechanisms of change. Not unlike qualitative work, such
29 approaches also generate interesting, and at times counter-intuitive, findings about
30 neighbourhood change (see, for example, Freeman *et al.*'s 2015 conclusion that there is no
31 elevated mobility out of those London neighbourhoods experiencing gentrification).
32
33
34
35
36
37
38

39 However, in contrast to the quasi-experimental approach of Freeman *et al.* (2015) which
40 said little about future trends, this paper has used innovative ML techniques to highlight
41 neighbourhoods that are likely to significantly improve or decline by 2021. As well as
42 noting the residualisation of some parts of outer London, our results suggest continuing
43 'uplift' in Inner East London and the spread of this process to the Outer Boroughs. Changes
44 in neighbourhood status are, not unsurprisingly, strongly associated with house prices, the
45 proportion of males and females in work for more than 30 hours a week, household
46 incomes, and the share of knowledge workers, homeworkers, and professionals. It is these
47 factors, as opposed to local amenities or travel, that appear worthy of more detailed
48 exploration. That said, recent political developments, such as Brexit and changes to
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 London's infrastructure (e.g. Crossrail), mean that, while the specific predictions in this
5 paper are unlikely to be accurate, they still provide a basis for further comparative
6 investigation.
7
8

9
10 As a demonstration of the capabilities of Machine Learning in an urban studies context, this
11 paper is a useful marker of the need for a rapprochement across the 'qualitative/quantitative
12 divide'. We are not claiming to have explained or 'solved' the problem of neighbourhood
13 change, nor are we suggesting that our approach supersedes the intensive, on-the-ground
14 work undertaken by so many before, but it does open a new 'front' in our attempts to
15 understand and, ultimately, anticipate neighbourhood transition. We hope that, in making
16 these predictions about change in London, we are ultimately able to identify the ways that
17 improvement or regeneration can occur without incurring displacement or disconcerting
18 social change. Perhaps our predictions will be wrong for all the right reasons?
19
20
21
22
23
24
25
26
27
28
29

30 **Acknowledgements**

31
32 First, we would like to acknowledge the valuable contribution of Dr. Elizabeth
33 Sklar, Jordan's co-supervisor on the original work that ultimately led to this article;
34 her advice was integral to this research, though any errors or omissions remain ours
35 alone. In addition, Jordan also wishes to acknowledge the contribution of Ivy Du to
36 this work. We also made extensive use of the contributions of the many developers
37 who have made possible Scikit-Learn 0.18 (Pedregosa *et al.* 2011) and Pandas
38 (McKinney 2010) under version 3.6 of the Python programming language. The
39 reproducible notebooks are made possible by Jupyter 4.1.0 (Kluyver 2016). The
40 maps were created in QGIS 2.18 (Quantum GIS Development Team 2017). Other
41 figures were produced in R using ggplot2 (Wickham 2009). All tools are available
42 as Free Open Source Software. The codebase, including installation and
43 configuration script for the required Python libraries, is available for download at:
44
45
46
47
48
49
50
51
52
53 <https://github.com/jreades/urb-studies-predicting-gentrification>.
54
55
56
57
58
59
60

Appendix (production notes)

Our approach made intensive use of the Scikit-Learn toolkit 0.18 (Pedregosa *et al.* 2011) and Pandas (McKinney 2010) under version 3.6 of the Python programming language. The reproducible notebooks are made possible by Jupyter 4.1.0 (Kluyver 2016). The maps were created in QGIS 2.18 (Quantum GIS Development Team 2017). Other figures were produced in R using ggplot2 (Wickham 2009). All tools are available as Free Open Source Software. The GitHub repository is available at [not included to retain anonymity of peer review].

References

- Anselin, L. (1995). Local Indicators of Spatial Association–LISA. *Geographical Analysis*, 27(2):93–115.
- Arribas-Bel, D., P. Nijkamp and H. Scholten (2011). Multidimensional urban sprawl in Europe: A self-organizing map approach. *Computers, Environment and Urban Systems*, 35(4):263–275.
- Arribas-Bel, D., J. Patino and J. Duque (2017). Remote sensing-based measurement of living environment deprivation: Improving classical approaches with machine learning. *PLoS ONE*, 12(5):e0176684.
- Atkinson, R. (2000). Measuring gentrification and displacement in Greater London. *Urban Studies*, 37(1), 149–165.
- Brunsdon, C. (2016). Quantitative methods I: Reproducible research and quantitative geography. *Progress in Human Geography*, 40(5):687–696.
- Barton, M. (2016). An exploration of the importance of the strategy used to identify gentrification. *Urban Studies*, 53(1):92–111.
- Benson, M. and E. Jackson (2017). Making the middle classes on shifting ground? Residential status, performativity and middle-class subjectivities in contemporary London *British Journal of Sociology* 68(2):215–233.

- 1
2
3
4 Bostic, R.W. and R.W. Martin (2003). Black home-owners as a gentrifying force?
5 neighbourhood dynamics in the context of minority home-ownership. *Urban Studies*,
6 40(12):2427–2449.
7
8
9
10 Boy, J.D. and J. Uitermark (2016). How to study the city on Instagram. *PloS*
11 *ONE*, 11(6):e0158161.
12
13
14 Breiman, L. (2001) Random forests. *Machine Learning*, 45(1):5–32.
15
16 Butler, T. and L. Lees (2006). Super-gentrification in Barnsbury, London: globalization and
17 gentrifying global elites at the neighbourhood level. *Transactions of the Institute of*
18 *British Geographers*, 31(4), 467–487.
19
20
21 Butler, T. and G. Robson (2001). Social capital, gentrification and neighbourhood change
22 in London: a comparison of three South London neighbourhoods. *Urban Studies*,
23 38(12):2145–2162.
24
25
26 Butler, T., C. Hamnett and M.J. Ramsden (2013). Gentrification, education and
27 exclusionary displacement in East London. *International Journal of Urban and*
28 *Regional Research*, 37(2), 556–575.
29
30
31 Chapple, K. (2009). *Mapping susceptibility to gentrification: The early warning toolkit*.
32 Technical report, Centre for Community Innovation, University of California
33 Berkeley, August 2009. URL:
34
35 <<https://communityinnovation.berkeley.edu/reports/Gentrification-Report.pdf>>.
36
37
38
39 Chapple, K. and M. Zuk (2016). Forewarned: The use of neighborhood early warning
40 systems for gentrification and displacement. *Cityscape: A Journal of Policy*
41 *Development and Research*, 18(3):109–130.
42
43
44 Clark, E. (1988). The rent gap and transformation of the built environment: Case studies in
45 Malmö 1860-1985. *Geografiska Annaler. Series B. Human Geography*, 241–254.
46
47
48 Cockings, S., A. Harfoot, D. Martin and D. Hornby (2011). Maintaining existing zoning
49 systems using automated zone-design techniques: methods for creating the 2011
50 Census output geographies for England and Wales. *Environment and Planning*
51 *A*, 43(10), 2399–2418.
52
53
54
55
56
57
58
59
60

- 1
2
3
4 Cole, H.V., M.G. Lamarca, J.J. Connolly and I. Anguelovski (2017). Are green cities
5 healthy and equitable? Unpacking the relationship between health, green space and
6 gentrification. *J Epidemiol Community Health*, 71(11), 1118–1121.
7
8
9 Dalton, C.M. and J. Thatcher (2015). Inflated granularity: Spatial ‘Big Data’ and
10 geodemographics. *Big Data & Society*, 2(2):1–15.
11
12 Davidson, M. and L. Lees (2005). New-build ‘gentrification’ and London’s riverside
13 renaissance. *Environment and Planning A*, 37(7):1165–1190.
14
15 Donaldson, D. and A. Storeygard (2016). The view from above: Applications of satellite
16 data in economics. *The Journal of Economic Perspectives*, 30(4):171–198.
17
18 Fan, C., S.J. Rey and S.W. Myint (2016, online early). Spatially filtered ridge regression
19 (SFRR): A regression framework to understanding impacts of land cover patterns on
20 urban climate. *Transactions in GIS*. doi: 10.1111/tgis.12240.
21
22
23 Foster, J. (1999). *Docklands: cultures in conflict, worlds in collision*. Psychology Press.
24
25 Freeman, L. (2005). Displacement or succession? Residential mobility in gentrifying
26 neighborhoods. *Urban Affairs Review*, 40(4):463–491.
27
28 Freeman, L. (2009). Neighbourhood diversity, metropolitan segregation and gentrification:
29 What are the links in the US? *Urban Studies*, 46(10):2079–2101.
30
31
32 Freeman, L., A. Cassola and T. Cai (2015). Displacement and gentrification in England and
33 Wales: A quasi-experimental approach. *Urban Studies* 53(13):2797–2814.
34
35 Gale, C.G. (2014). Creating an open geodemographic classification using the UK Census of
36 the Population. PhD thesis, University College London.
37
38 Gale, C.G. and P.A. Longley (2013). Temporal uncertainty in a small area open
39 geodemographic classification. *Transactions in GIS*, 17(4):563–588.
40
41 Galster, G. (2001). On the nature of neighbourhood. *Urban Studies*, 38(12):2111–2124.
42
43 Glass, R.L. (1964). *London: aspects of change*. MacGibbon & Kee.
44
45 Guerts, P., D. Ernst and L. Wehenkel (2016). Extremely randomized trees. *Machine*
46 *Learning*, 63(1):3-42.
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3
4 Hagenauer, J. and M. Helbich (2017). A comparative study of machine learning classifiers
5 for modeling travel mode choice. *Expert Systems with Applications*, 78:273–282.
6
7
8 Hamnett, C. (1983). Regional variations in house prices and house price inflation 1969–81.
9
10 *Area*, 15(2): 97–109.
11
12 Hamnett, C. (1984). Gentrification and residential location theory: a review and
13 assessment. *Geography and the Urban Environment: Progress in Research and*
14 *Applications*, 6:283–319.
15
16
17 Hamnett, C. (2003). Gentrification and the middle-class remaking of inner London, 1961–
18 2001. *Urban Studies*, 40(12):2401–2426.
19
20
21 Hamnett, C. (2009). Spatially displaced demand and the changing geography of house
22 prices in London, 1995–2006. *Housing Studies*, 24(3):301–320.
23
24
25 Hamnett, C. (2015). The changing occupational class composition of London. *City*, 19(2–
26 3):239–246.
27
28
29 Hamnett, C. and J. Reades (2018), Mind the Gap: implications of overseas investment for
30 regional house price divergence in Britain. *Housing Studies*,
31 doi: 10.1080/02673037.2018.1444151 .
32
33
34 Harris, A. (2012). Art and gentrification: pursuing the urban pastoral in Hoxton,
35 London. *Transactions of the Institute of British Geographers*, 37(2), 226–241.
36
37
38 Hochstenbach, C., S. Musterd and A. Teernstra (2015). Gentrification in Amsterdam:
39 Assessing the importance of context. *Population, Space and Place*, 21(8), 754–770.
40
41
42 Holland, M. (2012) Chatsworth Road: the frontier of Hackney’s gentrification, *The*
43 *Guardian*, URL: <[https://www.theguardian.com/uk/2012/jul/07/chatsworth-road-](https://www.theguardian.com/uk/2012/jul/07/chatsworth-road-frontline-hackney-gentrification)
44 [frontline-hackney-gentrification](https://www.theguardian.com/uk/2012/jul/07/chatsworth-road-frontline-hackney-gentrification)> [Last checked 17 February 2017]
45
46
47 Hristova, D., M. Williams, M. Musolesi, P. Panzarasa and C. Mascolo (2016) Measuring
48 urban social diversity using interconnected geo-social networks. International World
49 Wide Web Conference, Montreal, Canada, 2016; doi:10.1145/2872427.2883065.
50
51
52 Jackson, E. and M. Benson (2014). Neither ‘Deepest, Darkest Peckham’ nor ‘Run-of-the-
53 Mill’ East Dulwich: The Middle Classes and their ‘Others’ in an Inner-London
54
55
56
57
58
59
60

1
2
3
4 Neighbourhood. *International Journal of Urban and Regional Research*, 38(4),
5 1195–1210.

6
7
8 James, G., D. Witten, T. Hastie and R. Tibshirani (2013). *An introduction to statistical*
9 *learning*. London: Springer, 102, pp.303–368.

10
11 Karsten, L. (2003). Family gentrifiers: challenging the city as a place simultaneously to
12 build a career and to raise children. *Urban Studies* 40(12):2573–2584.

13
14
15 Kluyver, T., B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley,
16 J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing and
17 Jupyter Development Team (2016). Jupyter Notebooks—a publishing format for
18 reproducible computational workflows. *Positioning and Power in Academic*
19 *Publishing: Players, Agents and Agendas* (F. Loizides and B. Schmidt eds), IOS
20 Press. URL <<http://ebooks.iospress.nl/publication/42900>> [Last checked 20 August
21 2017].

22
23
24
25
26
27
28 Kolko, J. (2007). The determinants of gentrification. *SSRN*, December 2007. doi:
29 10.2139/ssrn.985714.

30
31
32 Lansley, G. (2016). Cars and socio-economics: understanding neighbourhood variations in
33 car characteristics from administrative data. *Regional Studies, Regional Science*,
34 3(1):264–285.

35
36
37 Lauria, M. and M.E. Stout (1995). The significance of scale in the analysis of
38 gentrification. *College of Urban and Public Affairs (CUPA) Working Papers*, 1991–
39 2000:9, University of New Orleans.

40
41
42 Leckie, G. (2009). The complexity of school and neighbourhood effects and movements of
43 pupils on school differences in models of educational achievement. *Journal of the*
44 *Royal Statistical Society: Series A (Statistics in Society)*, 172(3), 537–554.

45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
Lees, L. (2000). A reappraisal of gentrification: towards a ‘geography of gentrification’.
Progress in Human Geography, 24(3):389–408.

Lees, L. (2003). Super-gentrification: The case of Brooklyn Heights, New York City.
Urban Studies, 40(12):2487–2509.

- 1
2
3
4 Lees, L. (2014). The urban injustices of new Labour's "New Urban Renewal": The case of
5 the Aylesbury Estate in London. *Antipode*, 46(4), 921–947.
6
7
8 Ley, D. (1986). Alternative explanations for inner-city gentrification: a Canadian
9 assessment. *Annals of the Association of American Geographers*, 76(4), 521–535.
10
11 Li, Y. and Y. Xie (2018). A New Urban Typology Model Adapting Data Mining Analytics
12 to Examine Dominant Trajectories of Neighborhood Change: A Case of Metro
13 Detroit. *Annals of the American Association of Geographers*, doi:
14
15 10.1080/24694452.2018.1433016.
16
17
18 Liu, L., E.A. Silva, C. Wu and H. Wang (2017). A machine learning-based method for the
19 large-scale evaluation of the qualities of the urban environment. *Computers,*
20
21 *Environment and Urban Systems*, 65:113-125.
22
23
24 Lloyd, C.D., N. Bearman, G. Catney, A. Singleton and P. Williamson (2016) PopChange.
25 Liverpool: Centre for Spatial Demographics Research, University of Liverpool.
26 URL: <[https://www.liverpool.ac.uk/geography-and-](https://www.liverpool.ac.uk/geography-and-planning/research/popchange/introduction/)
27
28 [planning/research/popchange/introduction/](https://www.liverpool.ac.uk/geography-and-planning/research/popchange/introduction/)>
29
30
31 Manley, D. and R. Johnston (2014). London: A dividing city, 2001–11? *City*, 18(6):633–
32
33 643, 2014.
34
35 Manley, E., J. Addison and T. Cheng (2015). Shortest path or anchor-based route choice: a
36 large-scale empirical analysis of minicab routing in London. *Journal of Transport*
37
38 *Geography*, 43:123–139.
39
40
41 Martin, D., C. Lloyd and I. Shuttleworth (2011). Evaluation of gridded population models
42 using 2001 Northern Ireland Census data. *Environment & Planning A*, 43(8):1965–
43
44 1980.
45
46 Mavrommatis, G. (2011). Stories from Brixton: Gentrification and Different
47
48 Differences. *Sociological Research Online*, 16(2):1–10.
49
50 McKinney, W. (2010). Data Structures for Statistical Computing in Python, *Proceedings of*
51
52 *the 9th Python in Science Conference*, 51–56.
53
54
55
56
57
58
59

- 1
2
3
4 Melchert, D. and J.L. Naroff (1987). Central city revitalization: A predictive model. *Real*
5 *Estate Economics*, 15(1):664–683.
6
7
8 Meligrana, J. and A. Skaburskis (2005). Extent, location and profiles of continuing
9 gentrification in Canadian metropolitan areas, 1981-2001. *Urban Studies*,
10 42(9):1569–1592.
11
12
13 Minton, A. (2017) *Big Capital: who is London for?* Penguin.
14
15 Morenoff, J.D. and M. Tienda (1997). Underclass neighborhoods in temporal and
16 ecological perspective. *The Annals of the American Academy of Political and Social*
17 *Science*, 551(1):59–72.
18
19
20 Naik, N., S.D. Kominers, R. Raskar, E.L. Glaeser and C.A. Hidalgo (2017). Computer
21 vision uncovers predictors of physical urban change. *Proceedings of the National*
22 *Academy of Sciences*, 114(29):7571–7576.
23
24
25
26 Neal, S., G. Mohan, A. Cochrane and K. Bennett (2016). ‘You can’t move in Hackney
27 without bumping into an anthropologist’: why certain places attract research
28 attention. *Qualitative Research*, 16(5):491–507.
29
30
31
32 O’Sullivan, D. (2002). Toward micro-scale spatial modeling of gentrification. *Journal of*
33 *Geographical Systems*, 4(3), 251-274.
34
35
36 Office for National Statistics (nd). *Census geography*. URL:
37 <[https://www.ons.gov.uk/methodology/](https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography) geography/ukgeographies/censusgeography
38 - super-output-area-soa> [Last checked 17 August 2017].
39
40
41 Owens, A. (2012) Neighborhoods on the rise: A typology of neighborhoods experiencing
42 socioeconomic ascent. *City & Community*, 11(4):345–369.
43
44
45 Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P.
46 Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M.
47 Brucher, M. Perrot and E. Duchesnay (2011). Scikit-learn: Machine Learning in
48 Python, *JMLR* 12:2825–2830.
49
50
51
52 Quantum GIS Development Team (2017). Quantum GIS Geographic Information System.
53 *Open Source Geospatial Foundation Project*. URL: <<http://qgis.osgeo.org/>>.
54
55
56
57
58
59
60

- 1
2
3
4 Reades, J. and D. Smith. (2014) Mapping the ‘Space of Flows’: the geography of global
5 business telecommunications and employment specialisation in the London Mega-
6 City Region. *Regional Studies*, 48(1):105–126.
7
8
9
10 Reades, J. (2014) Mapping changes in the affordability of London with open-source
11 software and open data: 1997–2012. *Regional Studies, Regional Science*, 1(1):336–
12 338.
13
14
15 Redfern, P.A. (1997). A new look at gentrification: 2. A model of gentrification.
16 *Environment and Planning A*, 29(8):1335–1354.
17
18
19 Redfern, P.A. (2003). What makes gentrification ‘gentrification’? *Urban Studies*,
20 40(12):2351–2366.
21
22
23 Rey, S.J. (2014). Rank-based Markov chains for regional income distribution dynamics.
24 *Journal of Geographical Systems*, 16(2):115–137.
25
26
27 Santibanez, S.F., M. Kloft and T. Lakes (2015). Performance Analysis of Machine
28 Learning Algorithms for Regression of Spatial Variables: A Case Study in the Real
29 Estate Industry. Paper presented to *GeoComputation*, Dallas, pp.292–297.
30
31
32 Singleton, A.D., S. Spielman and C. Brunsdon (2016). Establishing a framework for Open
33 Geographic Information science. *International Journal of Geographical Information*
34 *Science*, 30(8):1507–1521.
35
36
37 Slater, T. (2009). Missing Marcuse: On gentrification and displacement. *City*, 13(2–3):292–
38 311.
39
40
41 Smith, N. (1996). *The new urban frontier: Gentrification and the revanchist city*.
42 Psychology Press.
43
44
45 Steif, K., A. Mallac, M. Fichman and S. Kassel (2017). *Predicting gentrification using*
46 *longitudinal census data*. URL: <[http://urbanspatialanalysis.com/portfolio/](http://urbanspatialanalysis.com/portfolio/predicting-gentrification-using-longitudinal-census-data/)
47 [predicting-gentrification-using-longitudinal-census-data/](http://urbanspatialanalysis.com/portfolio/predicting-gentrification-using-longitudinal-census-data/)> [Last checked 17 August
48 2017].
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3
4 Stevens, F.R., A.E. Gaughan, C. Linard and A.J. Tatem (2015). Disaggregating census data
5 for population mapping using random forests with remotely-sensed and ancillary
6 data. *PloS One*, 10(2):e0107042.
7
8
9
10 Sturgis, P., I. Brunton-Smith, J. Kuha and J. Jackson (2014). Ethnic diversity, segregation
11 and the social cohesion of neighbourhoods in London. *Ethnic and Racial*
12 *Studies*, 37(8):1286–1309.
13
14
15 Travers, T., S. Sims and N. Bosetti (2016). *Housing and inequality in London*. Technical
16 report, Centre for London.
17
18
19 van Criekingen, M. and J.-M. Decroly (2003). Revisiting the diversity of gentrification:
20 neighbourhood renewal processes in Brussels and Montreal. *Urban Studies*,
21 40(12):2451–2468.
22
23
24 van Ham M., D. Manley, N. Bailey, L. Simpson, D. Maclennan (2012) Neighbourhood
25 Effects Research: New Perspectives. In: van Ham M., Manley D., Bailey N.,
26 Simpson L., Maclennan D. (eds) *Neighbourhood Effects Research: New*
27 *Perspectives*, 1–21. Springer, Dordrecht.
28
29
30
31 Vickers, D. and P. Rees (2007). Creating the UK national statistics 2001 output area
32 classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*,
33 170(2):379–403.
34
35
36
37 Watt, P. (2008). ‘The only class in town? Gentrification and the middle-class colonization
38 of the city and the urban imagination’. *International Journal of Urban and Regional*
39 *Research*, 32:206–211.
40
41
42
43 Watt, P. (2013). ‘It's not for us’: Regeneration, the 2012 Olympics and the gentrification of
44 East London. *City*, 17(1):99–118.
45
46
47 Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New
48 York. URL: <<http://ggplot2.org/>>.
49
50
51 Wyly, E. (2014). Automated (post)positivism. *Urban Geography*, 35(5):669–690.
52
53
54 Xiao, N. (2016). Machine Learning in Richardson, D. (ed) *Encyclopaedia of Human*
55 *Geography*. John Wiley & Sons, Ltd, doi: 10.1002/9781118786352.wbieg0673.
56
57
58
59
60

1
2
3
4 Zhong, C., S.M. Arisona, X. Huang, M. Batty and G. Schmitt (2014). Detecting the
5 dynamics of urban structure through spatial network analysis. *International Journal*
6 *of Geographical Information Science*, 28(11):2178–2199.

7
8
9
10 Zukin, S., V. Trujillo, P. Frase, D. Jackson, T. Recuber and A. Walker (2009). New retail
11 capital and neighborhood change: boutiques and gentrification in New York City.
12 *City & Community*, 8(1):47–64.

13
14
15 Zukin, S., S. Lindeman and L. Hurson (2017) The omnivore’s neighborhood? Online
16 restaurant reviews, race, and gentrification. *Journal of Consumer Culture*, 17
17 (3):459–479.
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1. Model Comparison

Model	R²	Expl. Var.	MSE	MAE
Simple Linear Regression ¹	0.528	0.538	0.294	0.343
Multiple Linear Regression ²	0.639	0.640	0.225	0.305
Extremely Random Trees (Default)	0.649	0.653	0.219	0.284
Extremely Random Trees (Tuned)	0.699	0.703	0.188	0.259

¹ Using the strongest predictor variable (median house prices).

² Using all 166 variables.

Location of Detail within Decision Tree (roughly 75% Visible)

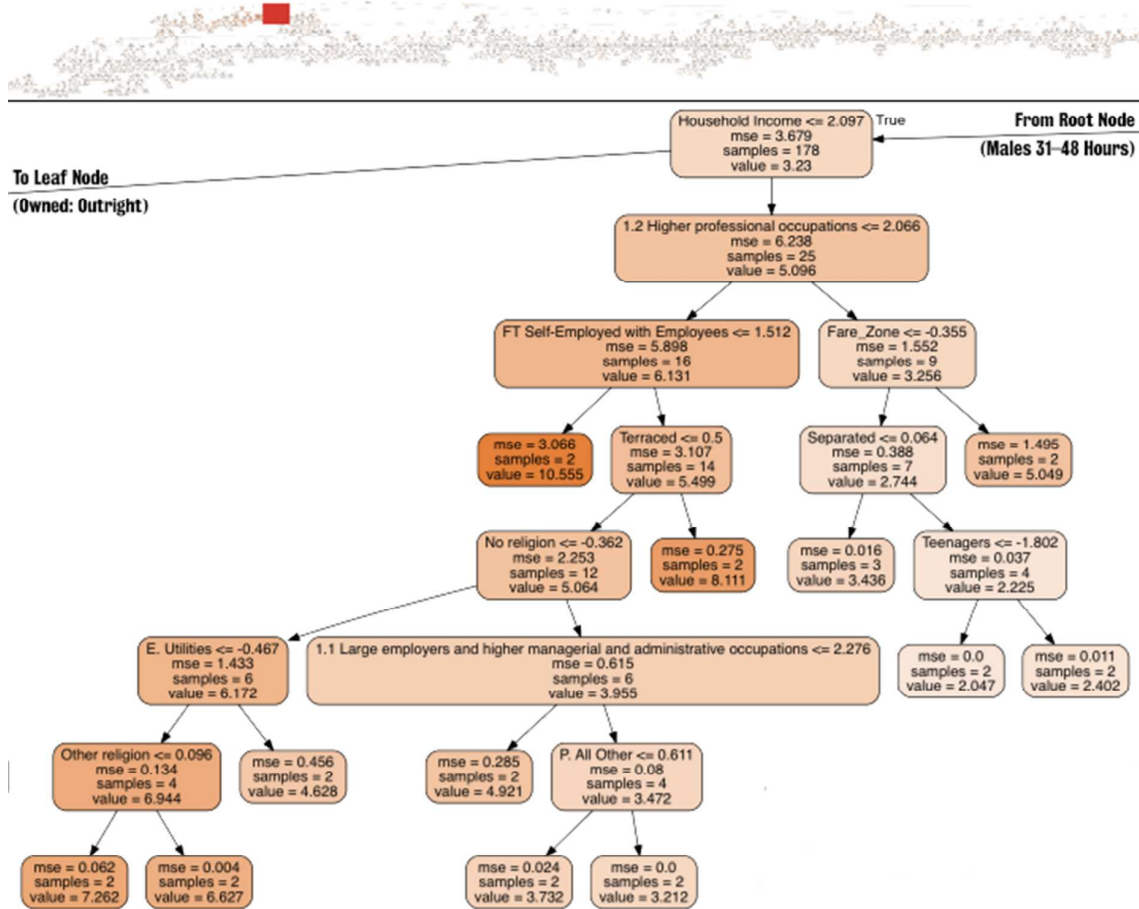


Figure 1. Detail from a regression tree used by the Random Forest in this research¹

¹ Each leaf node shows: the variable and value used in the split; the Mean Squared Error of the prediction for all observations in this region; the number of observations (samples); and the predicted value for observations in this region (this will usually be the mean).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

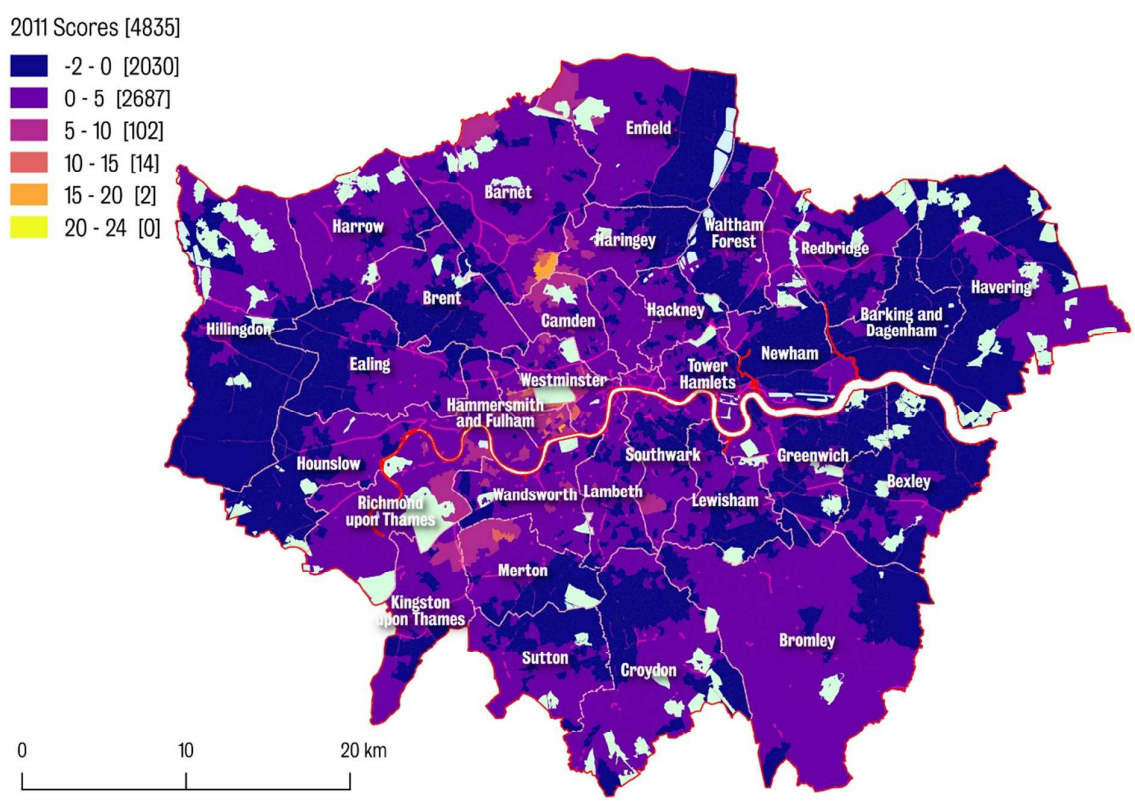


Figure 2. 2011 Status Scores for LSOAs

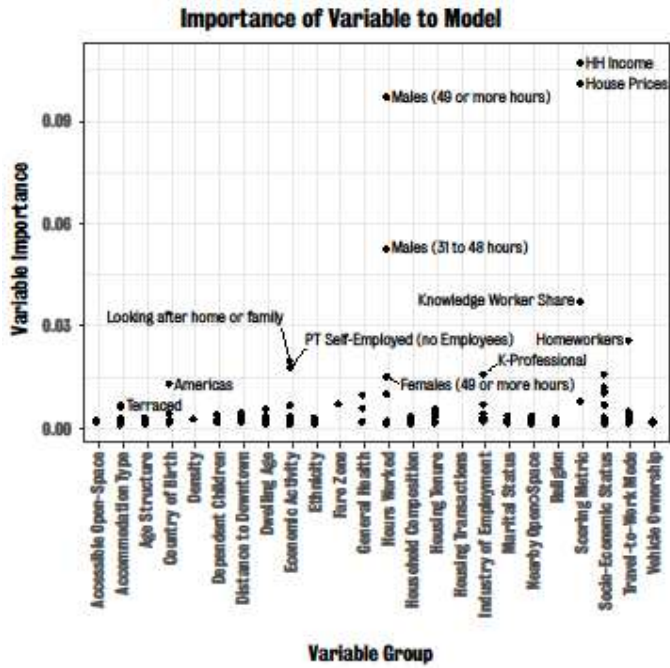


Figure 3. Parameter Importance to Tuned Model (Grouped by Variable Category)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

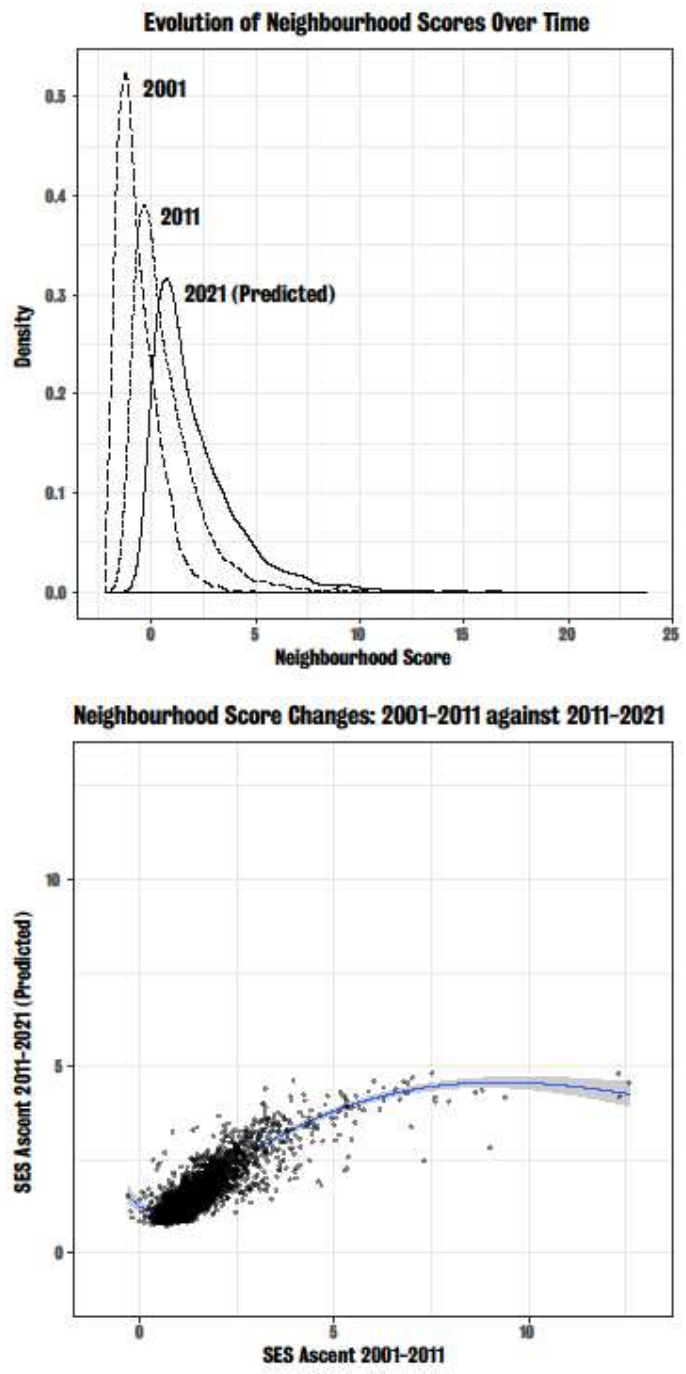


Figure 4a and b. Score Change Over Time

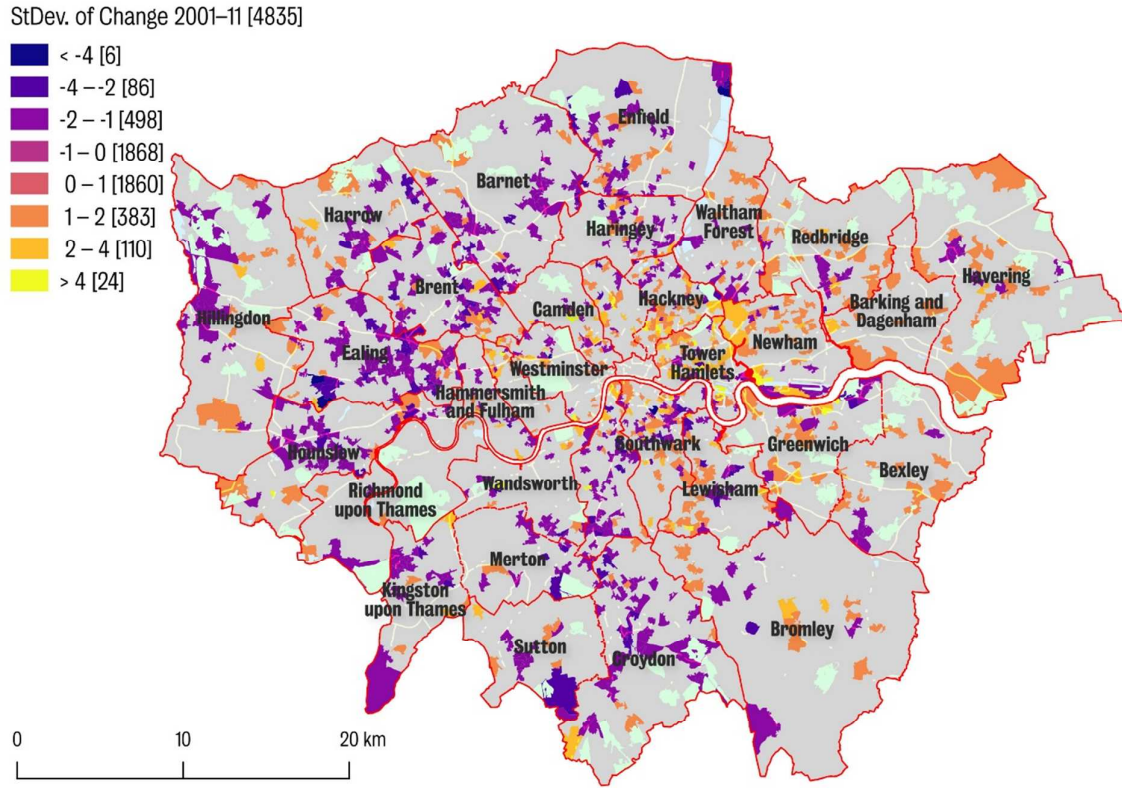


Figure 5. Standard Deviation of Change in Rank 2001–2011 (± 1 not shown)

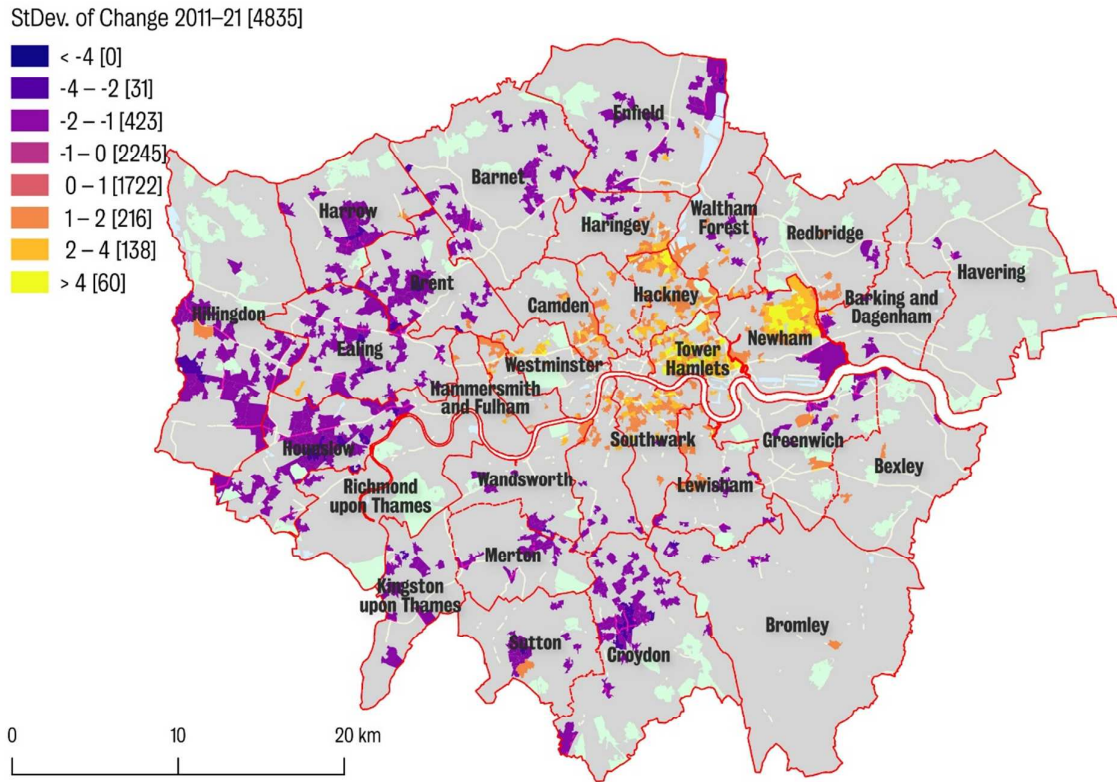


Figure 6. Standard Deviation of Change in Rank 2011–2021 (± 1 not shown)