urban science

MDPI

Article

Predicting Gentrification in England: A Data Primitive Approach

Jennie Gray [1,2,3,*], Lisa Buckner [4] and Alexis Comber [2,3]

[1] Geographic Data Science Lab, University of Liverpool, Liverpool L69 3BX, UK
[2] Leeds Institute for Data Analytics, University of Leeds, Leeds LS2 9JT, UK; a.comber@leeds.ac.uk
[3] School of Geography, University of Leeds, Leeds LS2 9JT, UK
[4] School of Sociology and Social Policy, University of Leeds, Leeds LS2 9JT, UK; l.j.buckner@leeds.ac.uk
* Correspondence: jhgray@liverpool.ac.uk

Abstract: Geodemographic classifications are useful tools for segmenting populations and have many applications but are not suitable for measuring neighbourhood change over time. There is a need for an approach that uses data of a higher spatiotemporal resolution to capture the fundamental dimensions of processes driving local changes. Data primitives are measures that capture the fundamental drivers of neighbourhood processes and therefore offer a suitable route. In this article, three types of gentrification are conceptualised, and four key data primitives are applied to capture them in a case study region in Yorkshire, England. These areas are visually validated according to their temporal properties to confirm the presence of gentrification and are then assigned to a high-level gentrification type. Ensemble modelling is then used to predict the presence, type, and temporal properties of gentrification across the rest of England. The results show an alignment of the spatial extent of gentrification types with previous gentrification studies throughout the country but may have made an overprediction in London. The periodicities of (1) residential, (2) rural, and (3) transport-led gentrification also vary throughout the country, but regardless of type, gentrification in areas within close proximity to one another have differing velocities such that they peak and complete within similar times. These temporal findings offer new, more timely tools for authorities in devising schedules of interventions and for understanding the intricacies of neighbourhood change.

Keywords: data primitives; neighbourhood change; gentrification; urban geography; urban dynamics

Citation: Gray, J.; Buckner, L.; Comber, A. Predicting Gentrification in England: A Data Primitive Approach. *Urban Sci.* **2023**, *7*, 64. https://doi.org/10.3390/urbansci7020064

Received: 4 December 2022
Revised: 6 May 2023
Accepted: 10 May 2023
Published: 13 June 2023

1. Introduction

Geodemographic classifications are useful tools for segmenting areas into groups or classes based on the socio-economic characteristics of their populations and sometimes of the areas themselves. They support applications in domains that seek to understand the spatial distribution of different neighbourhood types and the people they contain [1]. Geodemographic classifications are frequently constructed from population census data, which precludes the analysis of neighbourhood dynamics [2], although they have been used to infer neighbourhood change over decadal timespans [3]. The problem with using such classifications to understand dynamics is that the processes of interest may operate over varying spatial and temporal scales [4] that may not be captured by a decennial population census. There are consequently obvious limitations to classification-based approaches to quantifying neighbourhood-level processes through class allocation with temporally coarse data and the process of class allocation [5]. These are compounded by the assumption of synchronicity between process phase and measurement frequency [6], which is likely to be unmet.

A related issue is that classification is Boolean and allocates areas to the class (statistical cluster) to which they are closest in a multivariate feature space. This limits analysis to only dramatic changes in neighbourhood composition [5] and prevents nuanced analysis of geodemographic change. For example, depending on an area's position within the feature space (i.e., near the cluster centre or edge), different magnitudes of change are required

for class reallocation [6], with areas closer to the cluster edge requiring less change for reallocation than those near to the cluster centre. Additionally, within-cluster changes are not captured, even though they may indicate changes in cluster condition and quality or may be a signal of greater changes to come [7].

This paper adopts a data primitive approach to capture neighbourhood dynamics. The concept of data primitives [8] originated for land cover/land use mapping as a way of overcoming inconsistencies between different land use classifications in remote sensing [9] and has been used to link and separate land cover/land use semantics [10–12]. This paper extends the concept of data primitives into both the urban geography and temporal domains in an attempt to capture the neighbourhood process dynamics offered by data of a higher spatiotemporal resolution for a small area, thus capturing the nuances and dynamics of processes driving local changes [11].

In this study, interannual changes in four data primitives are examined to identify small areas that have been subject to gentrification, which are then manually validated. Using a national case study, three machine learning models are applied to selected annual data for small areas over a 10-year period that have been pre-processed in the same way as the training dataset. The aim is to predict the spatial distribution and timing of different types of gentrification nationally.

2. Background

2.1. Data Primitives

The absence of a dynamic element in geodemographic classification is a particular problem when dealing with change, such as occurs when an area undergoes gentrification. Conceptualising the data primitives—and the associated derived variables—as a kind of gentrification "space", this research draws on the data primitive approach to conceptualise gentrification as a change in the position of a small area within that data space over different time periods. In a neighbourhood analysis, these changes in position in a multi-variate feature space could be used to infer the changes in character experienced by a neighbourhood over time, and examining such shifts could be used to infer neighbourhood dynamics, to quantify process cycles, and to potentially predict future states [13]. This approach is, of course, dependent on the variables that are selected to identify and characterise the particular processes under investigation and the core drivers that characterise their changes. Further, the shifts in an area's position in the feature space must be filtered to determine potentially meaningful changes.

The data primitive approach is augmented with a change vector analysis (CVA) as a way to develop a clearer understanding of neighbourhood trajectories over time, as research in remote sensing change analyses have shown that the angle and magnitude of such positional changes can be used to infer the nature of the change [12]. CVA [14] originates from the remote sensing community and is used to determine land cover changes from shifts in a pixel's position in a multi-variate feature space of remote sensing image bands [15]. The magnitude of change is the Euclidean distance (length) between positions in the feature space, and the angle is the direction of the shift. Conceptually, the angle (direction) can help to discriminate between different types of change or different drivers of change [16], whilst the magnitude can be useful for comparisons within and among those change types [17].

A CVA generates measures of the Euclidian distance and the angle between two locations, x_1 and x_2, in a multivariate feature space. The distance, D, is calculated as follows:

$$D = \sqrt{(x_1 - x_2)^2} \quad (1)$$

The angle between the points, θ, is calculated from the dot product of the vectors of x_1 and x_2 in the following way:

$$\theta = \cos^{-1}\left(\frac{x_1.x_2}{|x_1||x_2|}\right) \quad (2)$$

where $|x_1|$ and $|x_2|$ are the absolute values of the vectors.

In this way, a CVA summarises a change across the full dimensionality of the data and has been found to be robust with respect to the nature and number of dimensions in the feature space [17]. In neighbourhood analyses, a CVA's magnitude and direction can be extracted and explored alongside changes in neighbourhood primitives. In this study, a CVA was applied to the single time period that most strongly indicated the presence of gentrification (see detail in the Section 3).

2.2. Gentrification

In UK-based studies, gentrification is often conceptualised as a class-based phenomenon: a product of a society rooted in a class-based hierarchy, whereby new residents of a gentrifying neighbourhood are of a higher social status than those in a prior time period [18]. It is driven by the in-migration of middle-class people who are more educated and more likely to be in professional occupations than the current (lower or working class) resident population. This increase in professional occupations is therefore often used to quantify the gentrification process [19]. There are also other effects: house prices increase, as do other costs, as a result of the changing nature of the local services reflecting the changing tastes of the new population [18]. This prices out the incumbent population while preventing the in-migration of lower- or working-class people. A further consequence of this situation is the ethnic "bleaching" of neighbourhoods as ethnic minorities, who tend to reside within lower-income neighbourhoods [20], are displaced. The consequence of this in- and out-migration is residential mobility or churn (the proportion of households that change) in gentrifying neighbourhoods, and it has recently been considered an important characteristic of gentrification [21].

While not necessarily exhaustive of the forms that gentrification might take—others [22,23] have noted super and green gentrification, for instance—these four data primitive domains, (1) professional occupation, (2) house price, (3) Black and Asian ethnicities, and (4) neighbourhood churn, should be sufficient to capture the changes associated with the fundamental drivers of gentrification in the UK.

3. Methods

To apply the data primitive approach, annual data covering these four key neighbourhood characteristics were collected, and machine learning models were trained on manually validated observations of gentrification.

3.1. Data

The data collected for Lower Super Output Areas (LSOAs) in England for the period 2010–2019 included the average house price, the proportion of people in professional occupations, the proportion of households that changed, and the proportion of the population that was Black and Asian. LSOAs are often used for neighbourhood-level analyses in the UK as they have a consistent population (~1500 people; ~500 houses) and have been found to be robust for analysing neighbourhood effects [24]. Table 1 summarises the attributes used as data primitives. These were collected from a range of open and safeguarded sources from which safeguarded data are only available via a successful application. Note that the professional occupation data are only available for Middle Super Output Areas (MSOAs), which have ~7500 people and ~2500 houses; this was spatially interpolated to LSOAs using area-weighted interpolation.

Two datasets were obtained from the Consumer Data Research Centre (CDRC) [25]. Modelled ethnicity proportions are safeguarded data for the *Black and Asian Ethnicities* data primitive, whilst the *Residential Mobility* primitive contains open data describing neighbourhood churn. Both datasets are products derived from the Linked Consumer Registers, which link the open electoral register with consumer registers supplied by value-added resellers [26]. The *Professional Occupation* data primitive was created by aggregating a selection of industries subjectively considered more "professional", as listed by the UK

government. The data in the *House Price* primitive were similarly freely available from the UK government.

From these, a dataset of 60 attributes was derived for each LSOA neighbourhood observation in the following way:

1. The data primitives were rescaled using *z*-scores and for each pair of years, a change score was calculated from the sum of the absolute change in the four data primitive values (45 attributes).
2. The characteristics of potential gentrification cycles were determined by identifying the start and end years and duration, the year of peak gentrification, the start to peak and start to end durations, and the cumulative sum of the gentrification scores to the peak year. These were counted and then filtered where possible to identify established cycles of gentrification with the following characteristics: a minimum of 2 years to reach peak gentrification; a peak score >1 standard deviations, as in Reades et al. [27]; a cycle end date of 2014 or greater; and selection of the cycle with the largest cumulative gentrification score to the peak year (eight attributes).
3. From these start and end years, the change in each data primitive was determined, and the magnitude and direction from a CVA of these positions in a normalised multivariate feature space were calculated (seven attributes).
4. Finally, a set of descriptive variables was collated to aid in the separation of gentrification types. These described neighbourhood distances to transport links (railway station, tram stop, bus station, and motorway junction), the counts of the number of transport links within 1 mile, 2.5 miles, and 5 miles, the minimum distance to any transport, distances to blue space and green space, and the number of green space access points within 500 m. A neighbourhood rural/urban descriptor [28] was also extracted (15 additional attributes).

The final list of variables used can be found in the Supplementary Materials.

Table 1. The data primitives, their spatial resolution changes associated with gentrification, and the measurement unit.

Data Primitive	Resolution	Change	Unit
House Price	LSOA	Increase	GBP
Professional Occupation	MSOA	Increase	Proportion
Residential Mobility (Churn)	LSOA	Increase	Proportion
Black and Asian Ethnicities	LSOA	Decrease	Proportion

3.2. Ensemble Modelling

Ensemble learning refers to the combination of multiple models to enable a more robust prediction, often with greater predictive performance than single machine learning models [29]. Three ensemble models, the gradient boosting machine (GBM), extreme gradient boost (XGBoost) and bootstrap aggregation (or bagging) models, were trained and evaluated via their confusion matrices and sensitivity and specificity. GBM iteratively refines an initial model by examining the error within the previous model, improving upon weak learners until some accuracy or iteration threshold is reached [30]. XGBoost is like GBM but also includes regression penalties within the boosting equation, with regularization controlling overfitting and often generating better-performing models [31]. Bagging is based on the concept of model averaging; it differs from boosting by training single models in parallel, rather than iteratively, and averages them to yield more accurate predictions [32].

Several models were created to predict:

1. The presence of gentrification (binary: whether present or not, with responses of *None* or *Gentrification*);
2. The type of gentrification (with responses of *None*, *Residential*, *Rural*, and *Transport*);
3. The temporal properties associated with the predicted type of gentrification (start, peak, and end years).

The training dataset was split with a 70:30 train/test ratio using a bootstrap approach to ensure the response variable had the same distribution in the splits. Models for predicting the presence and type of gentrification were initialized with the neighbourhood characteristic variables, data primitives, change vectors, and the gentrification indicators over the 45 time periods throughout the study. The temporal properties were predicted with all the previous variables, the predicted gentrification type, and the additional temporal variables. The models were cross-validated with repeated k-fold cross validation and were hyperparameter-tuned to find the optimal parameters relevant to the specific model. Predictions were generated and evaluated against the test sample via model accuracy, kappa value, and confusion matrices. The best-performing models with respect to these metrics were chosen and then fit to the entire training set to create the final models for the prediction in England. The England dataset was created in the same way as the training dataset, using the same combination of variables. When predicting the temporal properties, the models were run as regressions and rounded to the nearest year. Prediction probabilities for the classifications (presence of gentrification; type of gentrification) were also retained, particularly for type since the characteristics of the types of gentrification can often overlap. The probabilities can provide an indication as to the likelihood that a neighbourhood will gentrify and the likelihood of the type of gentrification, highlighting confusion and where potential misclassification may occur.

3.3. Case Study and Training Data

This research is based on a case study of South Yorkshire, a metropolitan county in the north of England. It is a suitable training ground for developing a national model due to its variation in landscape, built-up areas, and subsequent mixes of land use and neighbourhood types. The west is distinguished by the Peak District National Park, and the region sits upon the Yorkshire Coalfield, which is home to many quarries, industrial areas, mines, and mining villages. There are urban and rural settlements, large cities, farming communities, and commuting towns by different modes. The case study therefore covers a range of neighbourhood types, though it is landlocked and not comprehensive in its coverage of neighbourhood types.

The training dataset consisted of 853 LSOAs. Change vectors, which were created via a function that included modified code from the rastercva function of the RStoolbox R package [33], a range of neighbourhood characteristics, and some previously calculated indicators of change. These indicators represented change in relation to each time period between 2010–2019 (every year, every two years, every three years, and so on), resulting in 45 unique time periods with indicators of change. Within the dataset, there were 123 LSOAs with an associated cycle of gentrification, all of which were visually validated via Google Earth and Google Street View [34], a method gaining in popularity (see [35–38] for example). According to a neighbourhood's data primitives, its characteristics, and visual observation, it was allocated to one of three broad gentrification types: residential, rural, or transport gentrification. Three of these 123 LSOAs were classified as none, due to a lack of visual evidence of gentrification and limited changes observed within the data; 60 were classified as residential, 20 were classified as rural, and 40 were classified as transport.

4. Results

To recap, a dataset of 79 attributes was derived, 60 of which were derived from the 4 data primitives, and 15 of which were taken from contextual features. These attributes were used to train three ensemble models for South Yorkshire, and the results were vali-

dated manually. The best-performing model was then retrained for England as a whole. Bivariate models were used to predict the presence of gentrification, multivariate models were used to predict the type of gentrification, and finally, regression models were used to predict the temporal properties of the predicted types of gentrification.

The first models were trained and fit to predict the presence of gentrification, with a binary response of gentrification or no gentrification. Table 2 shows that when fit on training data for South Yorkshire, bagging outperformed GBM and XGBoost, with accuracy and kappa values of 99.65 and 0.985, respectively. Two Type 1 errors were present, with 2 None LSOAs predicted as gentrification. This represents a sensitivity of 1 and a specificity of 0.997. The bagging model was then fit to predict gentrification in England, resulting in 4556 LSOAs, around 14% of the LSOAs in England, predicted to have experienced gentrification throughout the 2010–2019 study period.

Table 2. Model results for predicting binary gentrification in South Yorkshire.

Model	Accuracy (%)	Kappa
GBM	98.94	0.957
Tree Bagging	99.77	0.985
Linear XGBoost	99.30	0.971

Figure 1 shows that the results of the tree bagging model: neighbourhoods predicted to have gentrified are scattered throughout the country, from major cities such as London, Manchester, and Leeds to the more rural inlands between these major urban areas. See Figure 2 for a reference map of these built-up areas.

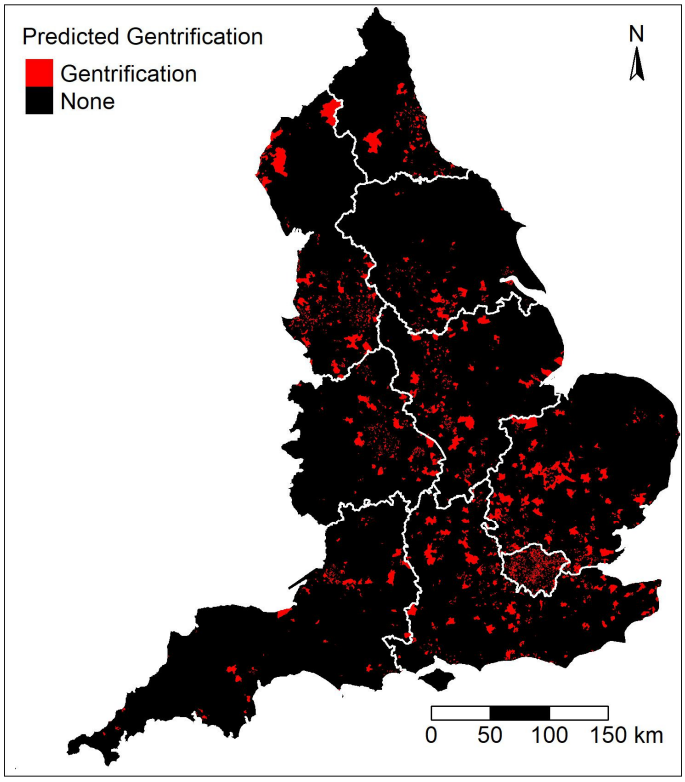

Figure 1. Probabilities for the binary prediction of the presence of gentrification in England.

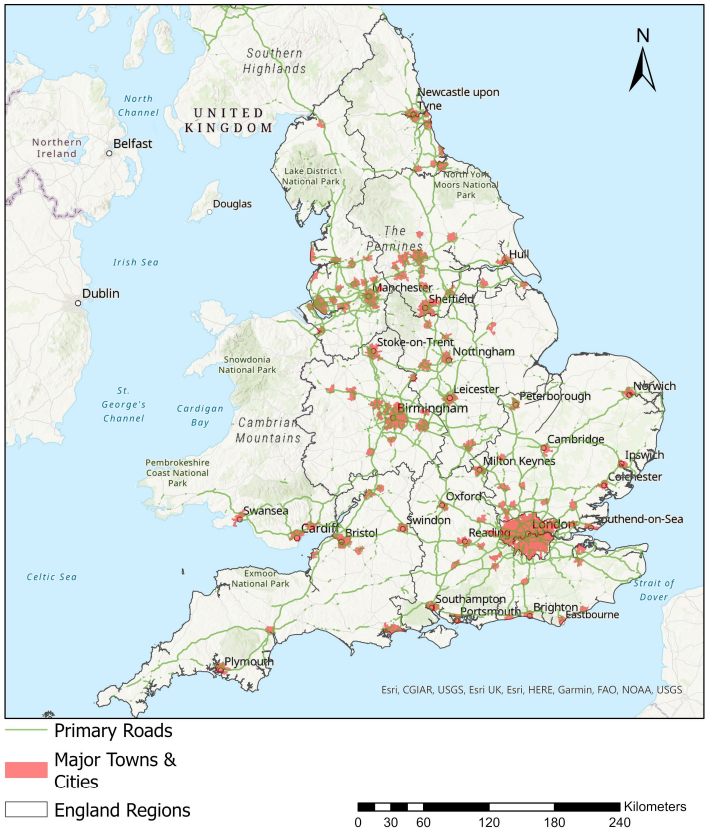

Figure 2. Reference map of England with major road networks and built-up areas.

The next models were the multivariate models, which were used to predict the type of gentrification, with responses of none, residential gentrification, rural gentrification, or transport gentrification. Table 3 shows that XGBoost outperformed GBM and bagging, with accuracy and kappa values of 98.59% and 0.945, respectively. Table 4 shows the confusion matrix, displaying the reference and predicted types of gentrification when applied to the whole of the training data. There was one misclassification for none, again a Type 1 error, which suggests that the non-gentrifying areas are sufficiently different from all types of gentrification in South Yorkshire but can confuse non-gentrifying with transport gentrification. Residential, rural, and transport gentrification all had Type 1 and Type 2 errors, with sensitivity values (true positives) of 0.95, 0.85, and 0.875, respectively. Though residential gentrification had the greatest sensitivity, it also had the most confusion and misclassification, with the lowest specificity value of 0.9917.

Table 3. Model results for predicting multivariate gentrification in South Yorkshire.

Model	Accuracy (%)	Kappa
GBM	98.01	0.922
Tree Bagging	98.48	0.941
Linear XGBoost	98.59	0.945

Figure 3 displays the probabilities of the different types of gentrification at the national level, displaying the presence of overlaps between residential and transport gentrification. Bardaka et al. [39] found that transit increases property values in neighbourhoods up to one mile from a station, which could explain some of the confusion between residential and transit gentrification. Figure 3d finally displays the overall predicted types of gentrification, a total of 4526 LSOAs, which is equivalent to 14% of the neighbourhoods in England.

Table 4. Confusion matrix of the GBM for predicting the type of gentrification on test data.

Reference Predicted	None	Residential	Rural	Transport
None	732	0	0	0
Residential	0	57	3	4
Rural	0	0	17	1
Transport	1	3	0	35

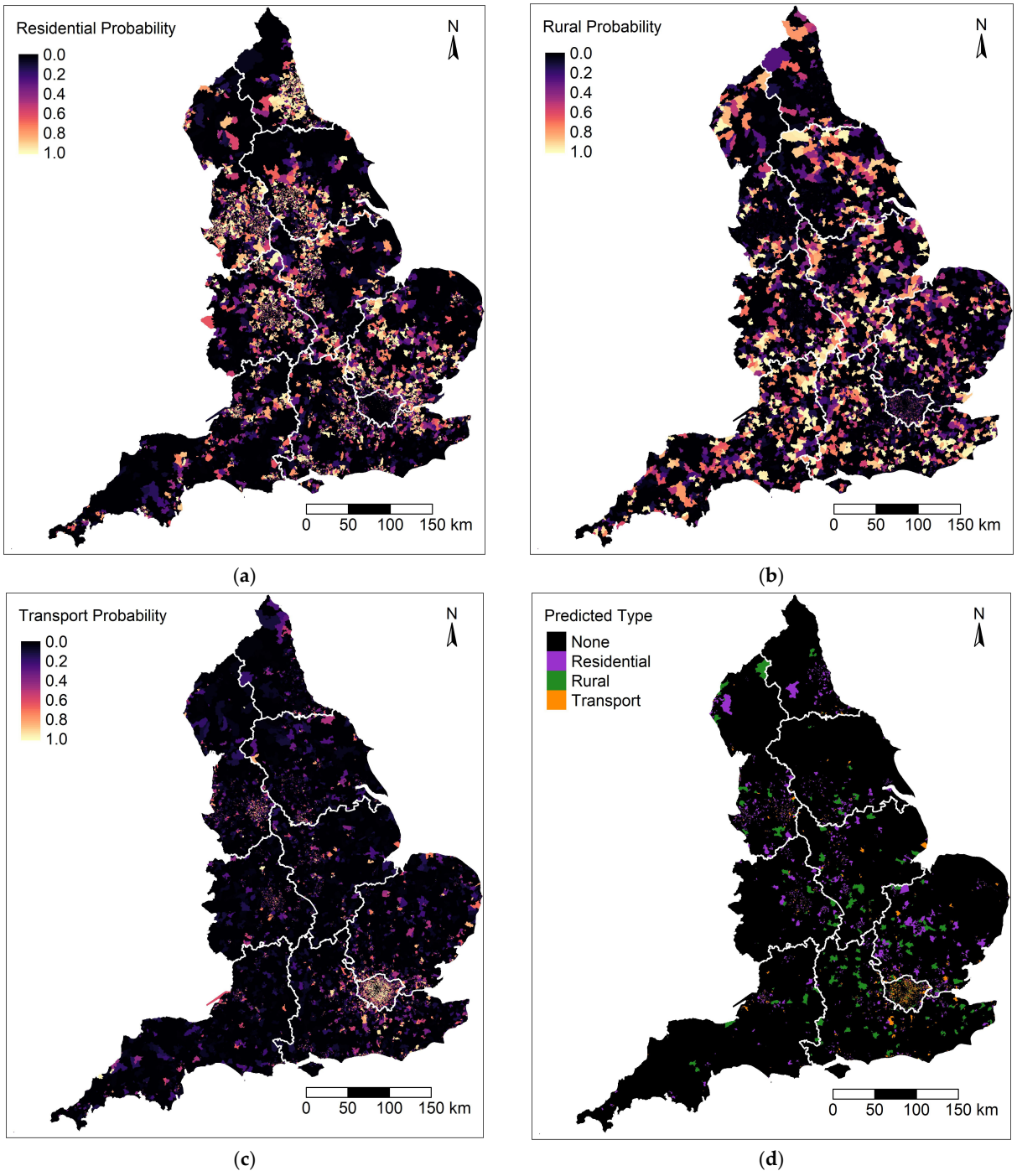

Figure 3. The probabilities of gentrification types: (**a**) residential gentrification probability; (**b**) rural gentrification probability; (**c**) transport gentrification probability; and (**d**) the overall predicted type of gentrification.

Residential gentrification (Figure 3a) was the most extensively predicted type of gentrification in England during the 2010–2019 study period and was predicted around major urban conurbations, including the outskirts of Greater London, Manchester, Newcastle, Birmingham, Nottingham, and Leeds. This supports previous research on gentrification within these cities: for example, gentrification in Newcastle was connected to development-driven (new-build) gentrification, a facet of residential gentrification [40]. State-led-replacement development-driven gentrification has also been experienced in Salford, Manchester, with negative impacts on those displaced [41].

The larger rural LSOAs distort the maps, but overall, rural gentrification (Figure 3b) is predicted with lower probabilities than residential gentrification. Rural gentrification in England between 2010–2019 occurred outside of major conurbations, often within proximity to national parks such as the North York Moors and Areas of Outstanding Natural Beauty. This highlights the pull of the rural idyl and supports previous research that explored rural gentrification in protected areas of England [42]. The residential and rural probability patterns are the inverse of one another.

Transport gentrification (Figure 3c) appears as the least likely type of gentrification and the most clustered; this is due to the densely populated LSOAs in which it was predicted. As is to be expected, transport gentrification was predicted around England's major transport hubs, such as London and Manchester. This supports previous research that found that the regeneration of a London Overground line catalysed gentrification [43]. Transport gentrification is also scattered in towns along major motorways running through the centre of England. Motorways contribute to suburbanization [44], which may facilitate gentrification in suburban neighbourhoods.

The final predicted gentrification types for England (Figure 3d) followed the highest probabilities for each gentrification type. Residential gentrification accounted for 54% (2454 LSOAs) and transport gentrification around 33% (1499 LSOAs), leaving rural gentrification with just under 13% (573 LSOAs).

The final models were run as regressions via XGBoost to predict the start, peak, and end years of the predicted gentrification cycles. These predictive models were applied to the LSOAs predicted with a gentrification type only (4526 LSOAs), opposed to the entire of England. Figure 4 shows the temporal predictions relating to the periodicity of gentrification: the start, peak, and end of gentrification in England. The gentrification start years were mostly predicted to be 2010 and 2011, but there were clusters with sequential starting years, mostly in the southern half of the country. The predicted peak years of gentrification indicate that clusters of LSOAs experiencing gentrification, regardless of their starting years, peaked at similar times, particularly in the south. Such clustering is also observed within the gentrification end years. This suggests that neighbouring localities of gentrification had varying velocities such that they peaked and completed their cycles at similar times. However, it does also show that although the model was applied to only those LSOAs that were predicted to gentrify, 141 LSOAs were consistently predicted without any temporal properties, suggesting no cycle of gentrification. However, the predicted zeros reflect areas where no temporal properties of the predicted gentrification were predicted.

The number of years taken to reach the peak of the process and the overall duration of the predicted gentrification in England were then calculated instead of being directly predicted. Table 5 shows the national averages of the duration, the number of years from the start to the peak, and the number of years from the peak to the end. Residential gentrification typically has a slower accumulation of change, taking longer to reach its peak before ending relatively swiftly, with the largest overall duration. On average, transport gentrification has similar manifestations to residential gentrification, with a more gradual accumulation of change, an accelerated peak to end, and a similar overall duration. Rural gentrification, however, has a more rapid accumulation of change, with a shorter start to peak duration before a relatively more gradual completion and a shorter average duration.

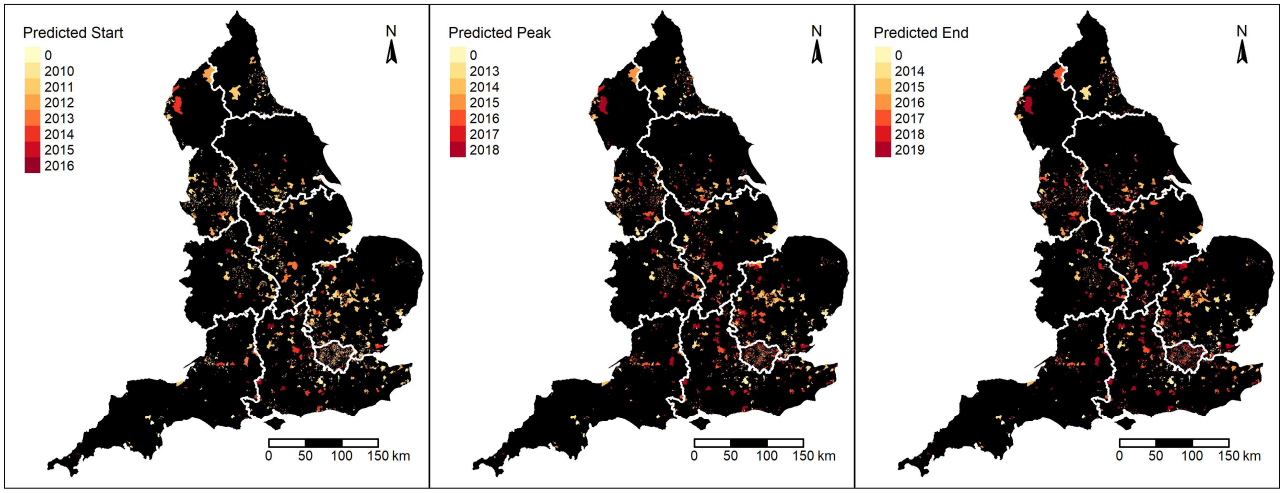

Figure 4. The predicted start, peak, and end years of gentrification in England.

Table 5. National averages of the duration, number of years from start to peak, and number of years from peak to end of the predicted gentrification in England.

Temporal Properties (Years)	Residential	Rural	Transport
Start to Peak	4.14	3.70	4.10
Peak to End	1.39	1.50	1.40
Duration	5.53	5.15	5.50

When observing these variables throughout space, there appear to be some more regional patterns, as shown in Figure 5, which demonstrate the duration of the predicted cycles of gentrification within England, faceted by region.

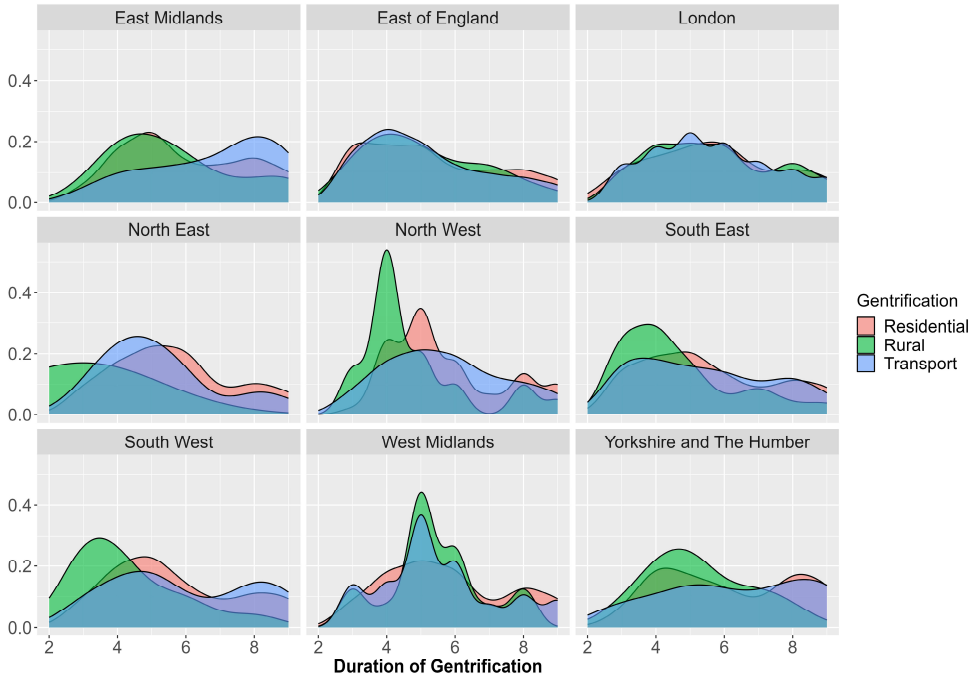

Figure 5. Density plots of the duration of predicted gentrification types by region.

The southeast and southwest had similar averages, with rural gentrification peaking at three years and residential and rural gentrification peaking at five and eight years, respectively. There is little differentiation between the different manifestations of gentrification

types within the east of England and again within London, with each gentrification type having a wide range of durations. The rural gentrification of Yorkshire and the Humber was predicted to have mid-length cycles, peaking between three and four years, whilst its transport gentrification was predicted to have longer cycles of around eight years. This contrasts with the northwest, where rural gentrification had a considerable peak at four years, with transport gentrification peaking at five years and residential gentrification having a wider range of durations.

These results therefore suggest that cycles of gentrification are not consistent throughout the country, and they have regional patterns that could be explored in greater depth.

5. Discussion

This research demonstrates that the data primitive approach is a viable alternative to and advancement upon traditional approaches to analysing neighbourhood change. Gentrifying neighbourhoods, as well as different types of gentrifying neighbourhoods, can be distinguished through the use of data primitives at a resolution of years, not decades. Predictive models can distinguish between gentrifying and non-gentrifying areas with a kappa of 0.99 (99% accuracy) and between different types of gentrification with a kappa of 0.95 (98.6% accuracy). Thus, gentrifying and non-gentrifying neighbourhoods and different types of gentrifying neighbourhoods are markedly different within their neighbourhood characteristics and composition of data primitive changes over time in England.

Much of the gentrification predicted between 2010 and 2019 aligned with previous studies, such as the residential gentrification predicted in Newcastle [40] and Manchester [41], the London Overground line transport gentrification in London [43], and the rural gentrification in Areas of Outstanding Natural Beauty such as the Cotswolds [42]. When comparing the London-based results of this study to [27], there are overlaps in areas predicted as gentrifying, suggesting that some of the gentrification in London is likely to have been experienced between 2010 and 2019, a time period that they could, however, only speculate for in [27].

However, contrasting predictions were also observed for some areas (e.g., [27] predicts decline where these results predict gentrification), suggesting opportunities for further investigation: it could be that our selected training region of South Yorkshire is unsuitable for predicting changes across all of England, but it is also just as likely that the additional temporal resolution of our data yields more timely predictions than ones derived from the Census.

The confusion presented between these outputs and those within the initial misclassification on training data could suggest that further separation between the types of gentrification is needed to generate more accurate predictions. However, it could also be that the gentrification types were too broad, and that more specific types of gentrification would have provided better separation. Nevertheless, the conceptualisations provided within this study demonstrate the value of adopting a data primitive machine-learning-based approach to predicting process-associated neighbourhood change.

The binary and the multivariate predictive models generated generally consistent figures, with around 14% of neighbourhoods predicted to have experienced gentrification throughout the study period, which also aligns with the number of LSOAs identified as gentrifying in the case study region (14%).

This research also demonstrates that data primitives can predict the temporal properties of predicted gentrification, providing the power to suggest the process phase of gentrification. These results are novel to this approach, afforded by the spatiotemporal resolution of the data primitives. Results suggest that there is no singular pattern of periodicity for residential, rural, or transport gentrification throughout England. However, when observing the overall duration by gentrification type, rural gentrification has the shortest overall predictions on average and transport gentrification the longest. This could potentially be because rural neighbourhoods are less dense and require less change to make significant impacts and are thus completed more rapidly. Alternatively, their true start date

may have been masked by the temporal boundary of the study, suggesting a synchronicity issue between the data and the phenomenon [6]. The length of transport gentrification could be explained by the investments that transportation brings [45] and their expanding catchments over time extending the length of the process [46].

Predictions of the peak and end of the gentrification cycle suggest that LSOAs experiencing gentrification within proximity to one another are likely to have differing velocities such that they peak and complete in similar time frames, aligning with the previous research [16]. A more in-depth exploration into the velocities of cycles via the interannual change vectors is warranted and is an interesting prospect of future work. However, presently, the greatest value of these novel process phase results is how they can be used. They offer great potential for planners and policy makers in developing a schedule of policy-based interventions, both to enhance the benefits of gentrification and to mitigate the consequences, such as displacement. This is because with a data primitive and machine learning approach, local authorities have the capability to predict whether a neighbourhood will gentrify, the type of gentrification they are likely to experience, and the process phase, and thus the sequence in which they will gentrify. This allows for the timely mitigation of consequential impacts on communities, such as by adopting community empowerment strategies to improve social cohesion in residential gentrification; enhancing tenant protections to reduce the polarisation associated with rural gentrification; and policy interventions for affordable housing to mitigate increased property prices in areas of transport gentrification around transport links [47]. Consequently, data primitives can provide local authorities with a tool for designing appropriate policy interventions at appropriate time periods to reduce the negative social, economic, and cultural impacts upon gentrifying neighbourhoods.

Limitations

There were 141 LSOAs with a predicted gentrification type (3%) that did not have any predicted temporal properties, suggesting no cycles identified and highlighting some level of confusion or misclassification between models. Thus, further explorations are required to generate more accurate predictions of the temporal properties. This could be achieved via a more explicit use of change vectors.

Neighbourhood characteristics and vectors of change were used alongside data primitives to predict three different types of gentrification in England: residential, rural, and transport. These gentrification types are not exhaustive, rather, they represent the aggregate validated types of gentrification identified in the training data region.

The visual validation of the detected gentrification in South Yorkshire and the assignment of LSOAs to a type of gentrification provided as a sound basis for the prediction of gentrification in England. However, it is an extremely time-consuming approach, and imagery is not always aligned with the years of interest [34]. Furthermore, it is also still an inherently subjective method of validation, with some difficulties in assigning LSOAs to just one type of gentrification for prediction. Nonetheless, this method validated 120 of the 123 identified LSOAs as gentrifying, representing an initial accuracy of 98% at capturing cycles of gentrification. Had the training data region been any larger, such method may not be viable without a larger team with more time and resources. Moreover, had a different region been selected, a different range of gentrification types may have been identified and consequently predicted for England via the validation.

Data primitives rely upon adequate spatiotemporal resolution data to generate dynamic insights into a process phase, but they are restricted within their temporal boundaries and are not yet capable of longer-term analyses. Thus, the universality of the approach is limited to those with suitable data representative of the fundamental drivers of neighbourhood processes. As the ubiquity of spatiotemporal data increases, some data, such as administrative data, are likely to increase in resolution and availability. However, as individuals become more aware of digital privacy, some will exercise their right of removal

from the open register, which may impact the quality of products that rely on them, such as the CDRC data used within this research.

6. Conclusions and Future Work

There are several routes into areas of future work, some of which were described above. Change vectors were introduced as a component of the data primitive approach to represent an area's magnitude and direction of change in a multidimensional feature space. However, due to this paper's focus on prediction, they were not used to their full capacity: the deeper analysis of the change vectors, and their angles specifically, is a potential future area of work. Previous research has shown that the angle of change can reflect the type of change occurring [13] and consequently the drivers of gentrification [16]. Thus, a deeper analysis of interannual change vectors may generate deeper insight into the quantification of the process phase. Understanding the angles may also aid in improving the overall model precision and recall.

Finally, a more suitable predictive model may be one that explicitly considers spatiality, particularly when extending analyses to national studies. For example, the geographically weighted gradient boosting machine, which is built to improve the GBM via smoothing kernels to weight the loss function [48], may be an appropriate alternative. Nevertheless, this approach is novel in its way of generating a deeper understanding of the temporal manifestation of the different types of gentrification in England.

To conclude, neighbourhood change is dynamic and can often have a process phase that is shorter than the typical decennial intervals used in analyses, meaning that many cycles are missed. Our results show that data primitives can identify and distinguish gentrifying neighbourhoods from non-gentrifying and between different types of gentrification. Furthermore, the nature of data primitives enables the identification and prediction of the temporal properties of gentrification, providing the power to suggest the process phase of gentrification. Subsequently, such predictions can provide local authorities with the capability to schedule a timetable of appropriate policies and interventions to increase benefits and mitigate the consequences of specific types of gentrification. The distinct academic value of this approach is its ability to detect, analyse, and predict temporal properties of neighbourhood processes. More focused and specialised investigations into neighbourhood change via data primitives may therefore aid in the dissecting and understanding of the complexities of neighbourhood change.

Although the data primitive approach is in its infancy, it has started to highlight and unpack deeper understandings of the temporal properties of gentrification in England. It has created novel findings in an innovative manner, contributing both to the literature on gentrification and the neighbourhood change methodology. With further refinement, this approach has enormous potential for understanding the intricate spatiotemporal relationships between different types of neighbourhood processes and how they change throughout space and time.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/urbansci7020064/s1, Table S1: List of all variables used within the analysis, and their description. Figure S1: Interactive Map of Predicted Gentrification

Author Contributions: Conceptualization, J.G., L.B. and A.C.; methodology, J.G.; software, J.G. and A.C.; validation, J.G.; formal analysis, J.G.; investigation, J.G.; data curation, J.G.; writing—original draft preparation, J.G.; writing—review and editing, L.B. and A.C.; visualization, J.G.; supervision, L.B. and A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the ESRC Centre for Doctoral Training–Data Analytics and Society, grant number ES/P000401/1.

Data Availability Statement: Data are not available due to the use of safeguarded data that are integral to the study.

Acknowledgments: The data for this research were provided by the Consumer Data Research Centre, an ESRC Data Investment, under project ID CDRC 600-01, ES/L011840/1; ES/L011891/1.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Harris, R.; Sleight, P.; Webber, R. *Geodemographics, GIS and Neighbourhood Targeting*; John Wiley and Sons: Hoboken, NJ, USA, 2005; Volume 7.
2. Longley, P.A. Geodemographics and the practices of geographic information science. *Int. J. Geogr. Inf. Sci.* **2012**, *26*, 2227–2237. [CrossRef]
3. McLachlan, G.; Norman, P. Analysing Socio-Economic Change Using a Time Comparable Geodemographic Classification: England and Wales, 1991–2011. *Appl. Spat. Anal. Policy* **2020**, *14*, 89–111. [CrossRef]
4. An, L.; Tsou, M.-H.; Crook, S.E.S.; Chun, Y.; Spitzberg, B.; Gawron, J.M.; Gupta, D.K. Space–Time Analysis: Concepts, Quantitative Methods, and Future Directions. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 891–914. [CrossRef]
5. Reibel, M.; Regelson, M. Neighborhood Racial and Ethnic Change: The Time Dimension in Segregation. *Urban Geogr.* **2011**, *32*, 360–382. [CrossRef]
6. Comber, A.; Wulder, M. Considering spatiotemporal processes in big data analysis: Insights from remote sensing of land cover and land use. *Trans. GIS* **2019**, *23*, 879–891. [CrossRef]
7. Zhu, Z. Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 370–384. [CrossRef]
8. Comber, A.J. The separation of land cover from land use using data primitives. *J. Land Use Sci.* **2008**, *3*, 215–229. [CrossRef]
9. Wadsworth, R.A.; Comber, A.J.; Fisher, P.F. Probabilistic Latent Semantic Analysis as a potential method for inte-grating spatial data concepts. In Proceedings of the Colloquium for Andrew U. Frank's 60th Birthday 2008, Vienna, Austria, 27 April 2008; Department of Geoinformation and Cartography: Vienna, Austria, 2008.
10. Adnan, M.; A Longley, P.; Singleton, A.D.; Brunsdon, C. Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases. *Trans. GIS* **2010**, *14*, 283–297. [CrossRef]
11. Comber, A.; Kuhn, W. Fuzzy difference and data primitives: A transparent approach for supporting different definitions of forest in the context of REDD+. *Geogr. Helv.* **2018**, *73*, 151–163. [CrossRef]
12. Xu, R.; Lin, H.; Lü, Y.; Luo, Y.; Ren, Y.; Comber, A. A Modified Change Vector Approach for Quantifying Land Cover Change. *Remote Sens.* **2018**, *10*, 1578. [CrossRef]
13. Gray, J.; Buckner, L.; Comber, A. Extending Geodemographics Using Data Primitives: A Review and a Methodological Proposal. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 386. [CrossRef]
14. Bovolo, F.; Bruzzone, L. A Theoretical Framework for Unsupervised Change Detection Based on Change Vector Analysis in the Polar Domain. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 218–236. [CrossRef]
15. Lindsay, J. Change Vector Analysis. 2012. Available online: https://jblindsay.github.io/ (accessed on 25 February 2023).
16. Gray, J.; Buckner, L.; Comber, A. Identifying Neighbourhood Change Using a Data Primitive Approach: The Example of Gentrification. *Appl. Spat. Anal. Policy* **2023**, 1–25. [CrossRef]
17. Johnson, R.D.; Kasischke, E.S. Change vector analysis: A technique for the multispectral monitoring of land cover and condition. *Int. J. Remote Sens.* **1998**, *19*, 411–426. [CrossRef]
18. Lees, L.; Slater, T.; Wyly, E.K. *The Gentrification Reader*; Routledge: London, UK, 2010; Volume 1.
19. van Ham, M.; Uesugi, M.; Tammaru, T.; Manley, D.; Janssen, H. Changing occupational structures and residential segregation in New York, London and Tokyo. *Nat. Hum. Behav.* **2020**, *4*, 1124–1134. [CrossRef]
20. Huse, T. Gentrification and Ethnicity. In *Handbook of Gentrification Studies*; Edward Elgar Publishing: Cheltenham, UK, 2018; pp. 186–204.
21. Yee, J.; Dennett, A. Stratifying and predicting patterns of neighbourhood change and gentrification—An urban analytics approach. *Trans. Inst. Br. Geogr.* **2022**, *47*, 770–790. [CrossRef]
22. Lees, L. Super-gentrification: The case of Brooklyn Heights, New York City. *Urban Stud.* **2003**, *40*, 2487–2509. [CrossRef]
23. Gould, K.; Lewis, T. *Green Gentrification: Urban Sustainability and the Struggle for Environmental Justice*; Routledge: London, UK, 2016.
24. Cockings, S.; Harfoot, A.; Martin, D.; Hornby, D. Maintaining Existing Zoning Systems Using Automated Zone-Design Techniques: Methods for Creating the 2011 Census Output Geographies for England and Wales. *Environ. Plan. A Econ. Space* **2011**, *43*, 2399–2418. [CrossRef]
25. Vij, N. Introducing the Consumer Data Research Centre (CDRC). *J. Direct Data Digit. Mark. Pract.* **2016**, *17*, 232–235. [CrossRef]
26. Lansley, G.; Li, W.; Longley, P.A. Creating a linked consumer register for granular demographic analysis. *J. R. Stat. Soc. Ser. A (Stat. Soc.)* **2019**, *182*, 1587–1605. [CrossRef]
27. Reades, J.; De Souza, J.; Hubbard, P. Understanding urban gentrification through machine learning. *Urban Stud.* **2018**, *56*, 922–942. [CrossRef]
28. Bibby, P.; Shepherd, J. *Developing a New Classification of Urban and Rural Areas for Policy Purposes—The Methodology*; Defra: London, UK, 2004.

29. Wu, H.; Levinson, D. The ensemble approach to forecasting: A review and synthesis. *Transp. Res. Part C Emerg. Technol.* **2021**, *132*, 103357. [CrossRef]
30. Sagi, O.; Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
31. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016.
32. Lee, T.H.; Ullah, A.; Wang, R. Bootstrap aggregating and random forest. In *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*; Springer: Cham, Switzerland, 2020; pp. 389–429.
33. Leutner, B.; Horning, N.; Leutner, M.B. Package 'RStoolbox'. R Foundation for Statistical Computing, Version 0.1. 2017. Available online: https://cran.microsoft.com/snapshot/2017-09-17/web/packages/RStoolbox/index.html (accessed on 12 May 2023).
34. Ilic, L.; Sawada, M.; Zarzelli, A. Deep mapping gentrification in a large Canadian city using deep learning and Google Street View. *PLoS ONE* **2019**, *14*, e0212814. [CrossRef]
35. Thackway, W.; Ng, M.; Lee, C.-L.; Pettit, C. Implementing a deep-learning model using Google street view to combine social and physical indicators of gentrification. *Comput. Environ. Urban Syst.* **2023**, *102*, 101970. [CrossRef]
36. Huang, T.; Dai, T.; Wang, Z.; Yoon, H.; Sheng, H.; Ng, A.Y.; Rajagopal, R.; Hwang, J. Detecting Neighborhood Gentrification at Scale via Street-level Visual Data. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022.
37. Ravuri, E.D. A Google Street View analysis of gentrification: A case study of one census tract in Northside, Cincinnati, USA. *Geojournal* **2021**, *87*, 3043–3063. [CrossRef]
38. Dickinson, S.T. *Exploring Green Gentrification in Established Urban Parks: A Study of Philadelphia's Neighborhood Parks*; Temple University: Philadelphia, PA, USA, 2022.
39. Bardaka, E.; Delgado, M.S.; Florax, R.J. Causal identification of transit-induced gentrification and spatial spillover effects: The case of the Denver light rail. *J. Transp. Geogr.* **2018**, *71*, 15–31. [CrossRef]
40. Cameron, S. Gentrification, housing redifferentiation and urban regeneration: 'going for growth' in Newcastle upon Tyne. *Urban Stud.* **2003**, *40*, 2367–2382. [CrossRef]
41. Hincks, S. Deprived neighbourhoods in transition: Divergent pathways of change in the Greater Manchester city-region. *Urban Stud.* **2016**, *54*, 1038–1061. [CrossRef]
42. Méténier, M. Lutte environnementale dans le parc national de Dartmoor: (re) définition d'un territoire de nature protégée par la dynamique conflictuelle. In *L'Espace Politique. Revue en Ligne de Géographie Politique et de Géopolitique*; Université de Reims Champagne-Ardenne: Reims, France, 2019.
43. Lagadic, M. Along the London Overground: Transport Improvements, Gentrification, and Symbolic Ownership along London's Trendiest Line. *City Community* **2019**, *18*, 1003–1027. [CrossRef]
44. Rocha, B.T.; Melo, P.C.; Afonso, N.; Silva, J.A. *Motorways, Urban Growth, and Suburbanisation: Evidence from Three Decades of Motorway Construction in Portugal*; Universidade de Lisboa: Lisbon, Portugal, 2021.
45. Chava, J.; Newman, P.; Tiwari, R. Gentrification in new-build and old-build transit-oriented developments: The case of Bengaluru. *Urban Res. Pract.* **2018**, *12*, 247–263. [CrossRef]
46. Lin, J.-J.; Yang, S.-H. Proximity to metro stations and commercial gentrification. *Transp. Policy* **2019**, *77*, 79–89. [CrossRef]
47. Ghaffari, L.; Klein, J.-L.; Baudin, W.A. Toward a socially acceptable gentrification: A review of strategies and practices against displacement. *Geogr. Compass* **2017**, *12*, e12355. [CrossRef]
48. Zhan, Y.; Luo, Y.; Deng, X.; Chen, H.; Grieneisen, M.L.; Shen, X.; Zhu, L.; Zhang, M. Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm. *Atmos. Environ.* **2017**, *155*, 129–139. [CrossRef]