Identifying multi-variable change over time - the ‘spot-the-difference’ function

Introduction

Topics such as gentrification, segregation, and population change are well-covered through urban planning literature however many of these, particularly historical, descriptions fail to incorporate the power of modern GIS technology in their analysis (Hillier, 2010).

This article will cover a spatial function, spot-the-difference() , developed in R to support quantitative research in urban planning.

R is an open-source programming language and environment that enables statistical computing and the creation of graphics supported by a wide range of geospatial packages.

To demonstrate the capabilities of this function the topic of gentrification in London between 2001 and 2011 will be investigated.

Gentrification

Observational analysis of the alterations of the social structure and housing markets of areas in inner-city London led Glass (1964) to create the term ‘gentrification’, described as a phenomenon in which the social character of a district is transformed through the displacement of its working class inhabitants.

Further research has identified this change as not only of social character – but a unique combination of social, physical and economic factors.

Causes of gentrification have been attributed to a range of variables such as change in manufacturing to service-based industries (Ley, 1986), new types of inner-city middle class profession (Hammett, 2003) and areas experiencing a widening in potential versus actual land value (Smith, 1987).

Spot-the-difference() provides a tool to analyse weighted combinations of different variables and how they may have changed over time – suitable for identifying scenarios such as gentrification which have no clear single variable identifier.

Methodology

The function contains three main phases: retrieving user input, data manipulation / analysis and mapping / data outputs (Figure 1).

Data Sources & Inputs

Demonstrating this function, gentrification will be defined as a change in the attributes of social class, housing tenure and house price.

Social Class

Social class was identified through the National Socio-economic classification (NS-SeC) a primary social classification from the UK Census. This value is judged from the most representative household member’s occupation, combined with information about employment status and supervisory responsibility (ONS, 2010).

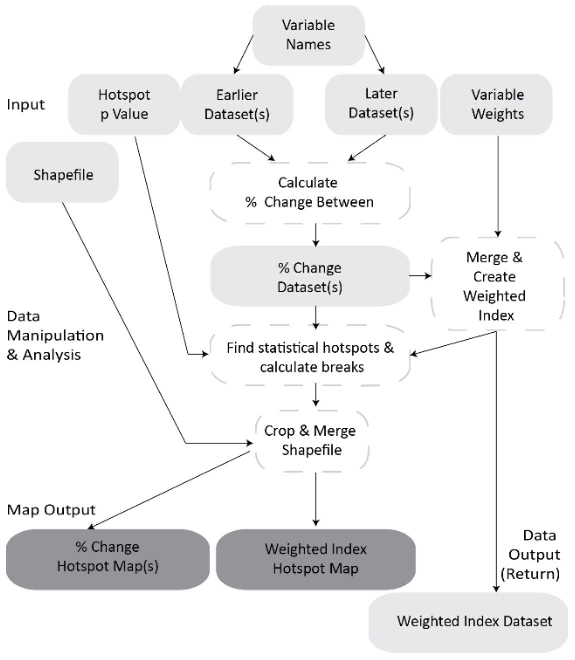


Figure 1. Flow chart illustrating process of spot-the-difference function. Annotated version (Appendix 1).

Tenure

Change in housing stock was identified through tenure data from the UK Census. This includes whether a household rents or owns accommodation and specifies whether this is privately or from other sources, such as social housing (ONS, 2013).

House Price

Residential property sales lodged at the Land Registry for England and Wales was utilised. This included properties, sale price and addresses which were reported to have sold in annual periods (Land Registry, 2013). These were aggregated into the same spatial units as compared datasets.

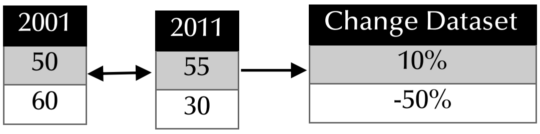
Scale

All datasets were prepared to be at the Lower-Super-Output Area (LSOA) scale for years 2001 and 2011. While only 2.5% of UK Output areas for 2011 differ from 2001 (UK Data Service, 2013), these were merged using lookup files to ensure consistency in the analysis (ONS, 2013).

Data Manipulation & Analysis

Percentage Change

Firstly, the function calculates the datasets into a percentage change dataset (Figure 2).



2001
50
60

2011
55
30

Change Dataset
10%
-50%

e.g. (2011 – 2001)/2001 * 100

Figure 2. Example percentage change tables.

Weighted index

Secondly, the percentage change datasets are combined. The individual variables are aggregated according to user-defined weightings (Figure 3) influenced by the weighted values used by the English Indices of Deprivation (Communities and Local Government, 2011).

Here these weightings partially reflect Atkinson (2000) who used number of professionals working as a proxy measure to identify areas of social evacuation.

% Change Managerial Professions	60%
% Change Private Rented Dwellings	35%
% Change Average Housing Price	5%

Figure 3. Weightings used for gentrification study

Hotspot calculation

The magnitude of disproportionality for each value within the datasets is calculated.

This is through dividing the difference between each value and the median by the difference between the hotspot cut-off and the median (Darrouzet-Nardi, 2013). A hotspot cut-off of 0.99 was chosen for this analysis.

Outputs

The following outputs were produced by the spot-the-difference() function.

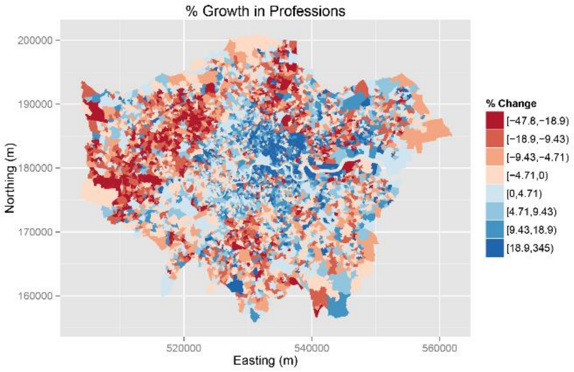


Figure 4. Example (1 of 3) percentage change output. Hotspots are highlighted in the colour extremes (darkest blue & red). Full output (Appendix 2).

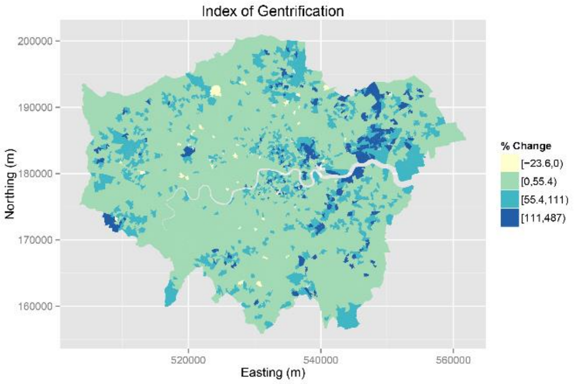


Figure 5. Example weighted index output – identifying areas that experienced a disproportionally high/low combination of weighted variables (classified as gentrification). Full output (Appendix 3).

Testing

A regression analysis found a moderate relationship (R-squared 37.2%) between the earlier dataset values and the data output – the weighted index (Appendix 5).

The coefficients indicate a positive correlation with price and stronger negative correlation for social status and privately rented dwellings.

This partially follows Smith (1987) arguing that gentrification can likely be the result of a wide gap between actual and potential land values.

Limitations

Relative versus absolute measures

This function originally included a comparison of an Index of Dissimilarity, a measure commonly associated with ethnic segregation (White, 1983), between high and low social status indicators.

Conflicts between relative versus absolute measures limited this usage. Potential sources, such as the English Indices of Deprivation, also face this issue (Communities and Local Government, 2011).

Modifiable areal unit problem

Flowerdew (2011) identified that the majority of 2001 Census data cases do not show significant differences at different scales, however variation in spatial units would need to be checked for variables used.

Displacement

An issue associated with analysing gentrification is identifying the flow of displacement (Atkinson, 2000). This analysis shows the change of variables that possibly resulted in population displacement, however is open to further extension to include this form of analysis.

Further Applications

Spot-the-difference() could assist with any spatial problem dealing with the difference between datasets using weighted variables – particularly those to do with population change over time.

Not limited to temporal data, the function could also be used in scenarios such as the comparison of datasets to calculate degrees of error between variables.

Further work incorporating its outputs into a regression model to make predictions of multi-variable phenomena would also be a useful adoption of this function.

References

Atkinson, R. (2011). Measuring gentrification and displacement in Greater London. *Urban Studies, 37*(1), 149-165.

Communities and Local Government. (2011). *The English Indices of Deprivation 2010.* Retrieved from GOV.UK: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/6320/1870718.pdf

Darrouzet-Nardi, A. (2013). *Package 'hotspots'.* Retrieved from R Project Packages: http://cran.r-project.org/web/packages/hotspots/hotspots.pdf

Flowerdew, R. (2011). How serious is the Modifiable Areal Unit problem for analysis of English census data? *Population Trends*, 106-118.

Glass, R. (1964). Introduction: aspects of change. In CenterForUrbanStudies (Ed.), *London: Aspects of Change* (p. xiii+xlii). London: MacGibbon and Kee.

Hillier, A. (2010). Invitation to mapping: how GIS can facilitate new discoveries in urban and planning history. *Journal of Planning History, 9*(2), 122 - 134.

Land Registry. (2013). *Price Paid Data.* Retrieved from Land Registry for England and Wales: http://www.landregistry.gov.uk/market-trend-data/public-data/price-paid-data

Ley, D. (1986). Alternative explanations for inner-city gentrification : A Canadian assessment. *Annals of the Association of American Geographers, 76*, 521-35.

ONS. (2010). *The National Statistics Socio-economic Classification.* Retrieved from Office for National Statistics: http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/soc2010-volume-3-ns-sec--rebased-on-soc2010--user-manual/index.html

ONS. (2013). *Geoportal Catalogue : Lookup files.* Retrieved from https://geoportal.statistics.gov.uk/geoportal/catalog/content/filelist.page

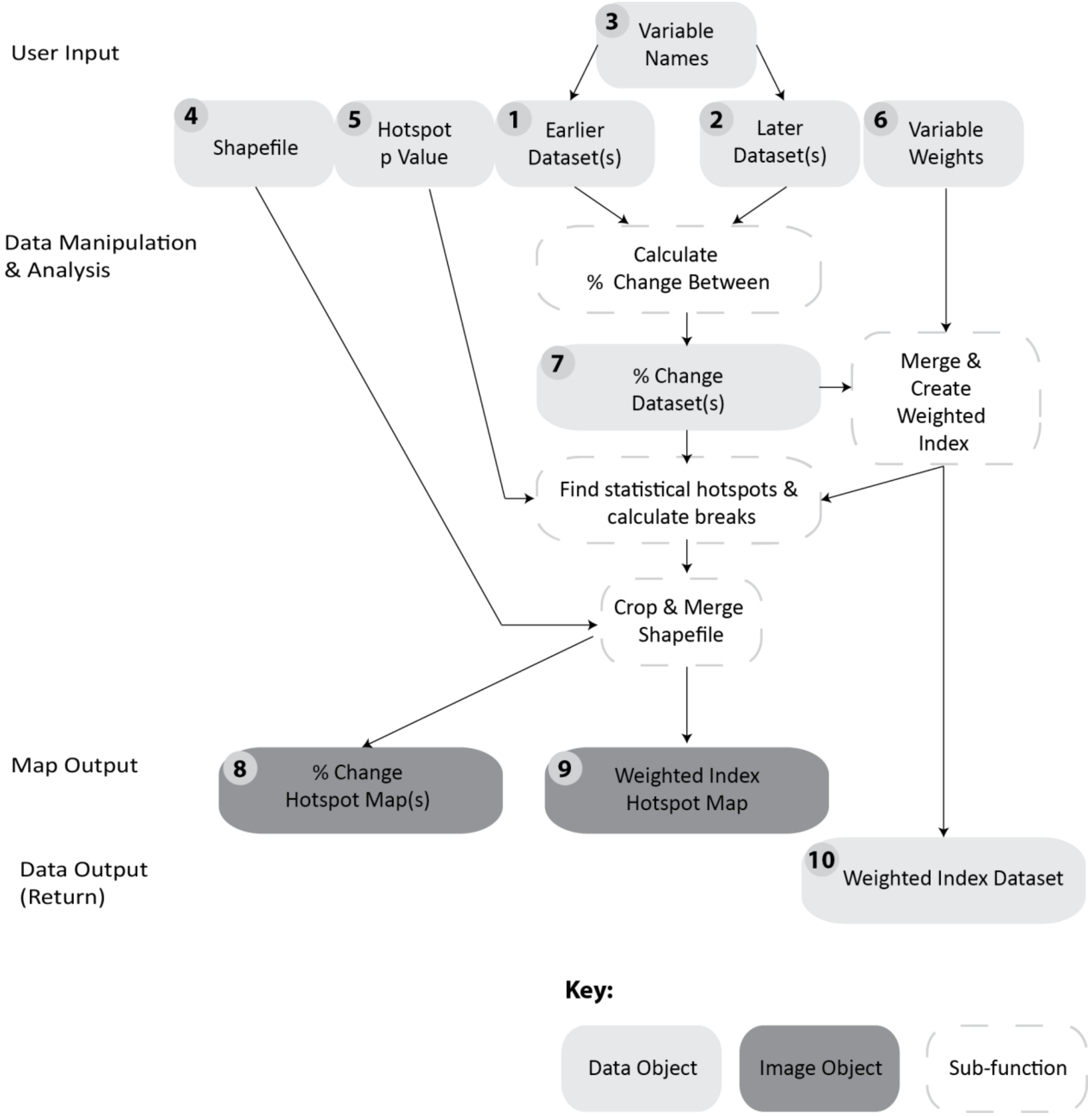
ONS. (2013). *Property By Tenure Type.* Retrieved from Office for National Statistics: http://www.ons.gov.uk/ons/taxonomy/index.html?nscl=Property+by+Tenure+Type

Smith, N. (1987). Gentrification and the rent gap. *Annals of the Association of American geographers*.

UK Data Service. (2013, June 23). *Comparing 2011 Census .* Retrieved December 2013, from http://ukdataservice.ac.uk/media/211697/5_compare_censuses.pdf

White, M. (1983). The measurement of spatial segregation. *American journal of sociology*, 1008-1018.

Appendix 1: Function Flow Chart



This shows main sub functions and data and image object(s) used and outputted in the spot-the-difference() function.

Brief descriptions of these used in the context of gentrification in London annotated:

(1) Land Registry Price Data 2001, NS-SeC 2001, Tenure 2001 - LSOA London

(2) Land Registry Price Data 2011, NS-SeC 2011, Tenure 2011 - LSOA London

(3) Average Residential Property Price, % LSOA in top 2 NS-SeC Classes, % Private Rented Accommodation

(4) Shapefile England and Wales LSOA Scale 2001

(5) P Value used : 0.99

(6) Social: 60%, Tenure : 35%, Price : 5%

(7) Datasets of the percentage change between 2001 and 2011 for variable (3) between (1) and (2)

(8) Three maps of percentage change for variables defined in (3) with highlighted hotspots

(9) A weighted index map of gentrification in London highlighted areas experiencing hotspots of this index value.

(10) The data output of this function to be used for use with other functions in analysing its value.

Appendix 2: Percentage Change Output Plots

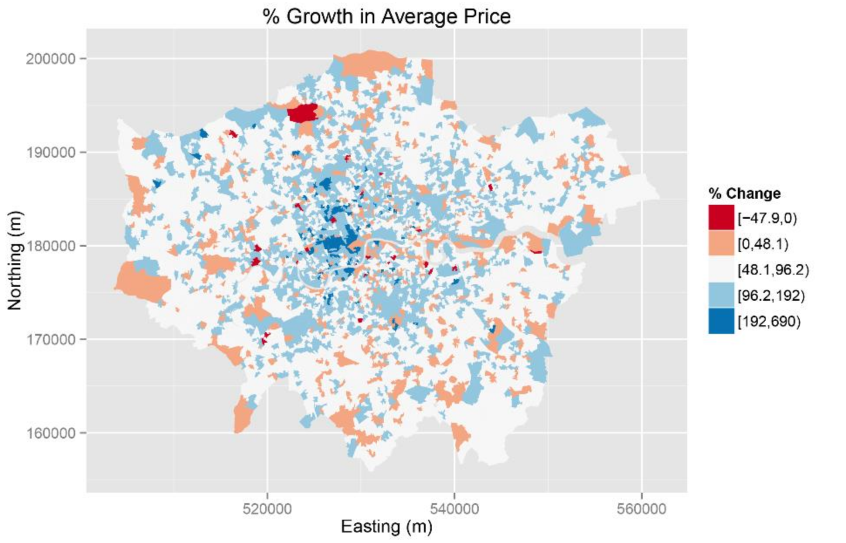


Figure A. Output map generated from specified 'House Price' dataset. This specifically shows change in average residential property price per Lower Super Output Area between the years of 2001 and 2011 (Land Registry, 2013). Positive hotspots (darkest blue) identified in inner-west London.

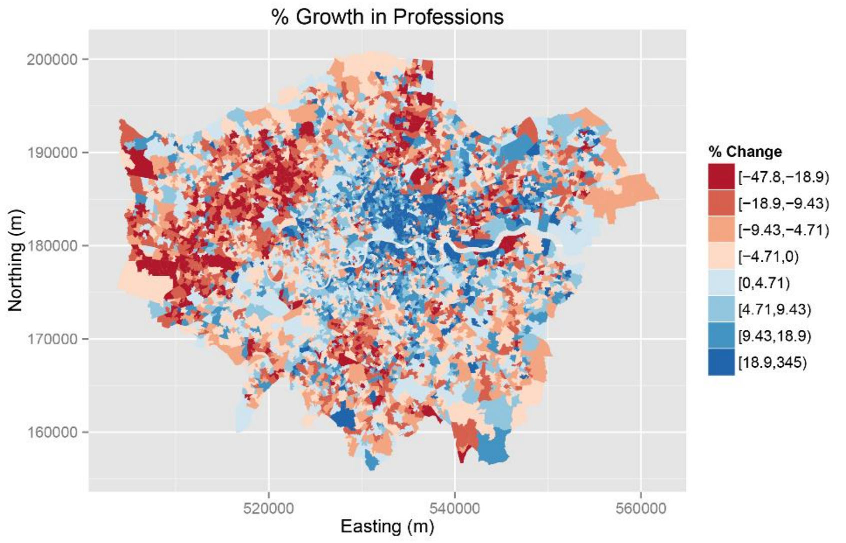


Figure B. Output map generated from specified 'Social Class' dataset. This specifically shows change in proportion of those in higher and lower 'managerial and professional' occupations' declared in the UK Census 2001 and 2011 (ONS, 2010). Positive hotspots (darkest blue) identified in inner-east London. Many negative hotspots (darkest red) also identified in outer-west.

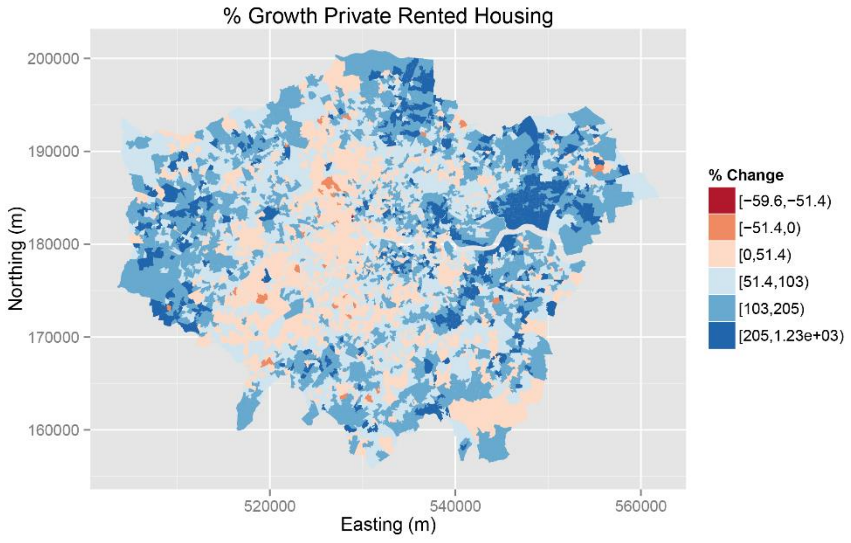


Figure C. Output map generated from specified 'Tenure' dataset. This specifically showing the change in percentage of household declared privately renting in the UK Census 2001 and 2011 (ONS, 2013). Positive hotspots (darkest blue) identified in inner to outer-east London as well as in outer north and west London.

Appendix 3: Gentrification Weighted Index Plot

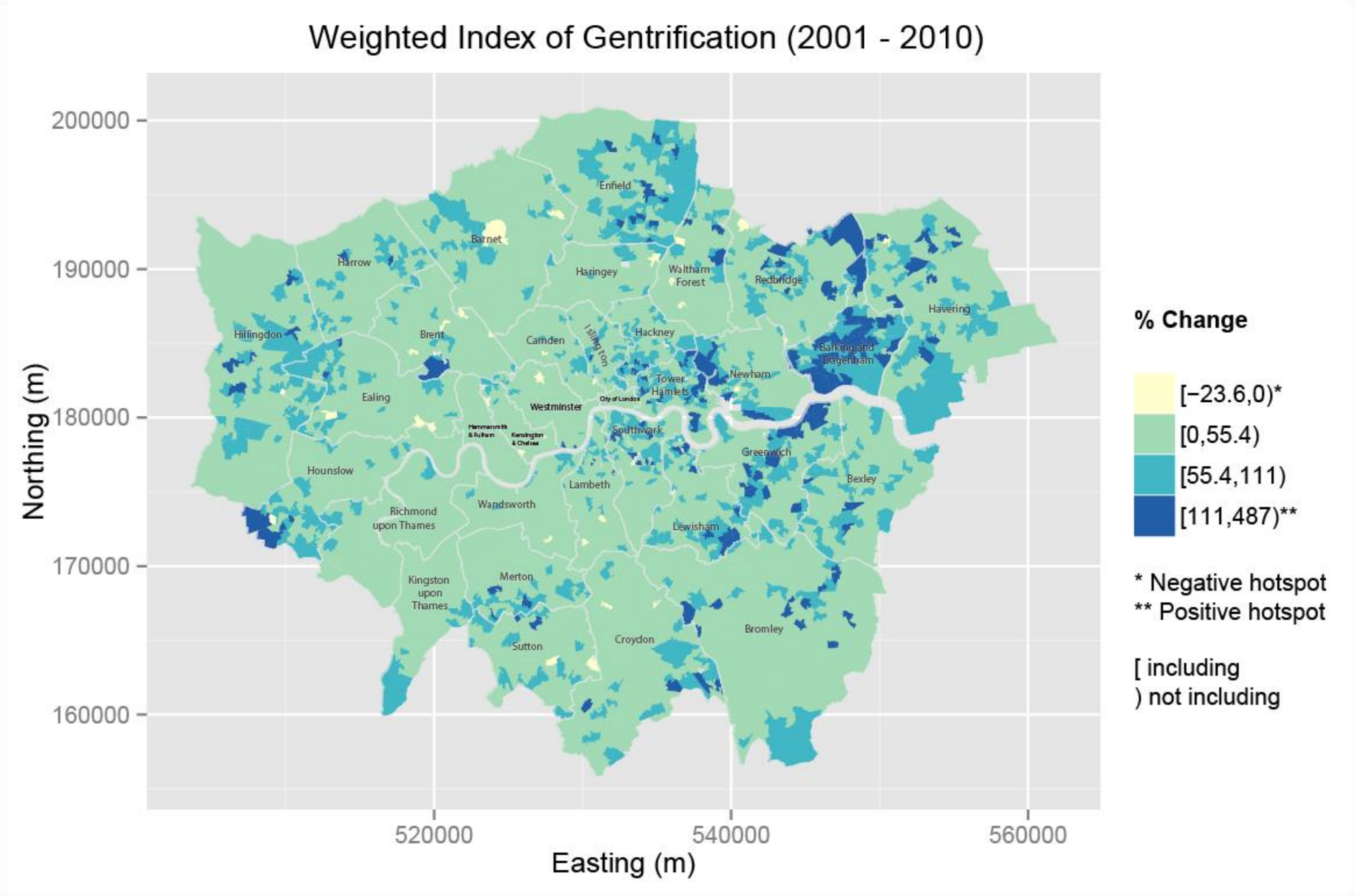


Figure D. This plot shows an example output of the weighted-index capabilities (annotated with borough names) from the spot-the-difference() function.

This plot identifies a number of clear areas of 'gentrification' (positive hotspot value) derived from change values of professional managerial class, privately-rented dwellings and household price (Appendix 2).

Appendix 4: User Documentation

Spot-the-difference() function

Introduction

This function takes a number of datasets from one year, and a number of datasets from another year and calculates how much specified variables have changed between them. The function also combines these datasets through the creation of a user-defined weighted index to identify multivariable phenomena.

The function's return value is an output of this weighted index which can then be used for testing such as regression analysis. During the running of the function the outputs are mapped and saved which identify hotspots of change within the individual variables as well as within the index.

Dependencies

This function depends on the following packages to run:

- rgdal
- maptools
- ggplot2
- RColorBrewer
- classInt
- hotspots

Information about these packages can be found here: http://cran.r-project.org/web/packages/available_packages_by_name.html

User Inputs

Due to the handling of many different datasets, the function also requires many user inputs, where *n* can be any number.

- A set of *n* datasets from one year. Stored in list variable (*earlier_datasets*) with *n* entries.

- A set of *n* datasets from later year. Stored in list variable (*later_datasets*) with *n* entries.

- The column headers of variables to be compared within each dataset. Stored in variable (*compared_fields*) with *n* entries.

- Titles for the map outputs. Stored in list variable (*map_titles*) with *n* entries. The main map title is stored under (main_map_title).

- Weightings for the weighted index. Stored in list variable (*weightings*) with *n* entries.

- A common identifier among each dataset. Stored in single string variable (*common_id*).

- The name of the shapefile. Stored in single string variable (*shapefile*), minus the file type (i.e. do not include '.shp').

- The id field of the shapefile. Stored in string variable (shapefile_id)

- The hotspot *p* value. Defaulted in the function as 0.99. Stored in numeric variable (hotspot_val).

- A number of other inputs have been marked within the function for modification:

Running the function

For ease of reading, the code of the function has been annotated where a hash and a question mark shows areas for user input (#?), a hash and an exclamation mark (#!) shows the beginning and end of the function. Outside of the main user variable declaration, within the function there are areas such as file names and map colours marked with (#?) that users can change).

Throughout running the function a number of values are printed out to assist in debugging should the values entered initially be incompatible with the function.

Using the function

Once the user inputs (#?) and the definition of the function has been run (between #! and #!) the function should be performed by this code. It is recommended storing it within a variable (e.g. A).

```
A<-spot_the_difference(earlier_datasets,later_datasets,compared_fields,common_id,shapefile,shapefile_id,map_titles,main_map_title,weightings,hotspot_val)
```

Function output

If the function has run successfully the an output similar to following should be visible in the R console:

```
[1] "Shapefile imported"
[1] "Diff. Calculated for Dataset # 1"
[1] "Calculating Hotspots for Dataset # 1"
[1] "Trying ggsave.."
[1] "Plot Saved"
[1] "Diff. Calculated for Dataset # 2"
[1] "Calculating Hotspots for Dataset # 2"
[1] "Trying ggsave.."
[1] "Plot Saved"
[1] "Trying ggsave.."
[1] "Plot Saved"
[1] "Merging Datasets"
[1] "Weighted Index Created"
[1] "Calculating Hotspots for WI"
[1] "Trying ggsave.."
[1] "Plot Saved"
[1] "Function Complete - Plots Saved &
Dataset Stored"
```

If the function is run by itself it will simply produce n+ 1 map outputs: *n* maps showing hotspots of percentage change between datasets and a map showing the hotspots of the weighted index of the combination of these maps.

If the function is stored as a variable this also produces the maps but the variable will contain the weighted index calculated by the function for use with other functions.

Appendix 5

Regression Results

```
> Rmodel <- lm(merged_newset$aggrecol ~ log10(merged_newset$avg_price) +
log10(merged_newset$perc_1_2) + log10(merged_newset$private_rented))

> summary(Rmodel)


Call:

lm(formula = merged_newset$aggrecol ~ log10(merged_newset$avg_price) +

   log10(merged_newset$perc_1_2) + log10(merged_newset$private_rented))


Residuals:

   Min     1Q Median     3Q    Max

-90.20 -14.14  -1.16   9.85 404.67


Coefficients:

                                    Estimate Std. Error t value Pr(>|t|)

(Intercept)                          118.523     12.244   9.680  < 2e-16 ***

log10(merged_newset$avg_price)        15.397      2.983   5.162 2.54e-07 ***

log10(merged_newset$perc_1_2)        -66.236      3.544 -18.691  < 2e-16 ***

log10(merged_newset$private_rented)  -50.818      1.284 -39.588  < 2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 28.66 on 4736 degrees of freedom

Multiple R-squared:  0.3732,    Adjusted R-squared:  0.3728

F-statistic: 939.8 on 3 and 4736 DF,  p-value: < 2.2e-16
```