Exploratory Spatial Data Analysis

Introduction

The use of spatial analysis and statistics is fundamental when conducting any form of GIS (Geographical Information Systems research. Spatial analysis involves the cohesion of two very individual forms of geographic information (Goodchild, 1986). One of which is the attribute of the spatial objects, the other in regards to the locational information of the spatial objects.

This study aims to investigate how different spatial analysis methods vary in appropriateness depending on the extent of the research area and consequently the size dataset used. In order to conduct this investigation, analysis of the ONS Land Registry's Price Paid data for 2016 is examined in context to both London and Hackney. The London extent gives a general overview of the dataset; whereas the Hackney borough provides more detailed analysis of the data (Figure 1). Two extents of analysis were performed due to singularities of viewing perspectives often restrict specific details of interest, or adversely obstruct universal trends and anomalies for the entire dataset (Butkiewicz et al., 2008). The study compares the appropriateness of simple spatial analysis method, the True Mean technique, and more complex interpolation methods such as Ordinary Kriging (OK) and Inverse Distance Weighting (IDW).

Figure 1: The study area of the project, created in ArcMap.

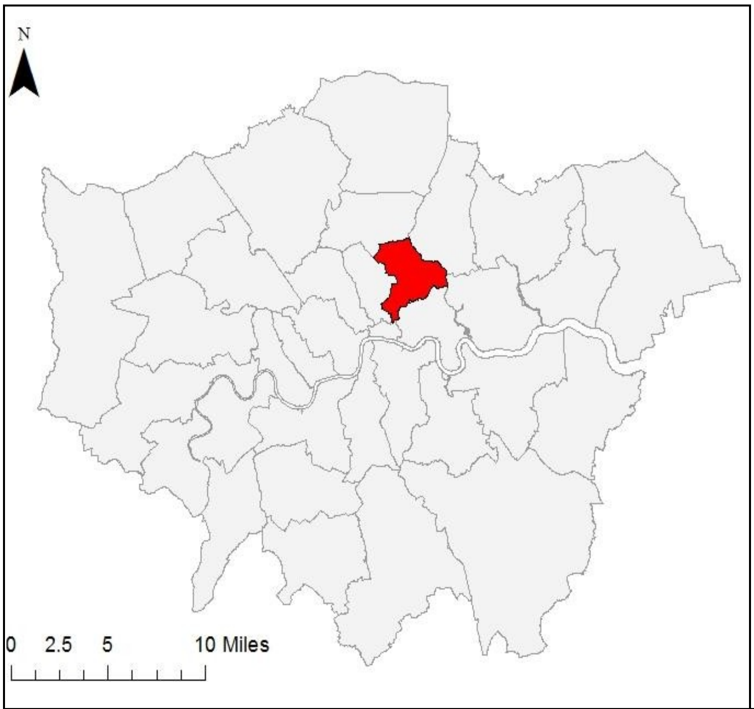


Figure 1a: Overview of the two study extents: London and Hackney. Polygon of London, with separating borders to show the different Borough Areas. Hackney highlighted in red.

Figure 1b: Enlarged map of Hackney (bordered in red), with the locations of the house price points associated within Hackney. The World Imagery Basemap used from ArcMap.

The spatial correlation between house prices is often associated to similar characteristics that nearby properties share within local markets (Yao and Fotheringham, 2015; Statch, 2009). In order to measure the spatial continuity of the dataset, Inverse Distance Weighting (IDW) and Original Kriging (OK) interpolation were performed. Spatial interpolation involves the procedure of forecasting the values of unsampled spatial sites using known attributes from observed points, based on the principle that similar attributes are correlated between points closer together (Xie et al., 2011; Schloeder et al., 2001).

Geostatistical methods use the spatial characteristics attributed to the data to analyse random events. These methods can be applied to interpret and predict spatial distribution of data based on approximate values in relation to the measurements calculated from other data points with known locations (Sarma, 2009). The study used the Original Kriging interpolation method, which identifies regularities throughout the spatial distribution of geospatial information by measuring the autocorrelation between point values (Ahsan and Parvez, 2014; Xie et al., 2011). (Cellmer, 2014)

Similarly with OK, IDW measures the linear combination of observed values with assigned weights (Xie et al., 2011). IDW interpolation is a geometric spatial analyst method that calculates the predicted values of points surrounding an individual point containing geospatial attributes, and outputs the interpolated points a weighting factored value based on the proximity of this point to the original data point to create a trend surface. The results are inversely proportional to the distance between two points, and therefore higher values to those that are closest. (Bhunia, Shit and Maiti; Xu, Guan and Zhou, 2015; Xie et al., 2011; Luo, Taylor and Parker, 2008).

Data

There were many underlying issues with the data provided, sourced from the ONS Land Registry's and Census statistics Price Paid data for 2016 which had to be 'cleaned' for spatial analysis. Much of the pre-processing 'cleaning' was undergone using the RStudio IDE due to its extensive ability to operate large datasets. The R-script for the pre-processing methods can be seen in the Appendix. The most necessary adjustment made to the dataset was to remove the extremities in house prices which weren't representative of a house price and therefore skewed the data. As can be seen in Table 1 & 2, demonstrating the highest and lowest 20 price values in the datasheet. Highlighted in red are examples of misrepresented data; Table 1, described as "Parking Spaces" equal to £200; Table 2, described as "Terminal 5" representing the Heathrow Airport terminal equal to £330000000. Both of these are examples of misrepresented data which should not be included within the dataset. Extremities such as these were removed by eliminating any data which lay outside the 2.5th and 97.5th percentiles.

Table 1: Table representing the lowest 20 lowest price values within the untouched London House Price dataset .

Postcode	X1.x	TransactionI[	Price	Date	PAON	SAON	Street	Locality	Town-City	District
HA03NG	284448	{2FD36065-2(	1	17/02/2016	191A	NA	EAST LANE	NA	WEMBLEY	BRENT
E143AJ	339728	{50F18103-C(	100	23/06/2016	73	NA	LOCKESFIELD	NA	LONDON	TOWER HAMLE
IG118ED	510785	{3E0330F0-B1	100	24/08/2016	15	NA	STATION PAF	NA	BARKING	BARKING AND [
N76JT	90060	{42A5A70A-0	100	10/06/2016	THE BEAUX ARTS BUILDING, 10 - 18	UNIT 11	MANOR GAR	NA	LONDON	ISLINGTON
N76JT	90063	{42A5A70A-0	100	10/06/2016	THE BEAUX ARTS BUILDING, 10 - 18	UNIT 16	MANOR GAR	NA	LONDON	ISLINGTON
N76JT	90056	{42A5A70A-0	100	10/06/2016	THE BEAUX ARTS BUILDING, 10 - 18	UNIT 7	MANOR GAR	NA	LONDON	ISLINGTON
N76JT	90058	{42A5A70A-0	100	10/06/2016	THE BEAUX ARTS BUILDING, 10 - 18	UNIT 9	MANOR GAR	NA	LONDON	ISLINGTON
N76JT	90057	{42A5A70A-0	100	10/06/2016	THE BEAUX ARTS BUILDING, 10 - 18	UNIT 8	MANOR GAR	NA	LONDON	ISLINGTON
N76JT	90051	{42A5A70A-0	100	10/06/2016	THE BEAUX ARTS BUILDING, 10 - 18	UNIT 1	MANOR GAR	NA	LONDON	ISLINGTON
RM64BE	619041	{47844C80-C	100	08/12/2016	9	NA	STATION RO/	CHADWELL H	ROMFORD	BARKING AND [
SE16BN	908116	{49B7852A-E	100	19/10/2016	SIGNAL BUILDING, 89 - 93	UNIT 2	NEWINGTON	NA	LONDON	SOUTHWARK
N76JT	90052	{42A5A70A-0	200	10/06/2016	THE BEAUX ARTS BUILDING, 10 - 18	UNIT 2	MANOR GAR	NA	LONDON	ISLINGTON
NA	324387	{47844C80-A	200	18/03/2016	PARKING SPACE 2 CLAYBURY HALL	NA	REGENTS DRI	NA	WOODFORD G	REDBRIDGE
NA	324388	{47844C80-A	200	18/03/2016	PARKING SPACE 4	NA	CLAYBURY H.	NA	WOODFORD G	REDBRIDGE
NA	324386	{47844C80-A	200	18/03/2016	PARKING SPACE 2	NA	CLAYBURY H.	NA	WOODFORD G	REDBRIDGE
NA	324397	{47844C80-A	200	18/03/2016	PARKING SPACE 3	NA	CLAYBURY H.	NA	WOODFORD G	REDBRIDGE
N76JT	90055	{42A5A70A-0	250	10/06/2016	THE BEAUX ARTS BUILDING, 10 - 18	UNIT 6	MANOR GAR	NA	LONDON	ISLINGTON
N98BU	93895	{404A5AF4-6	300	30/09/2016	1	NA	ST ALPHEGE	NA	LONDON	ENFIELD
E82NP	78130	{42A5A70A-0	455	04/11/2016	51 - 63	NA	RIDLEY ROAD	NA	LONDON	HACKNEY

Table 2: Table representing the 20 highest price values within the untouched London House Price dataset .

Postcode	X1.x	TransactionI[	Price	Date	PAON	SAON	Street	Locality	Town-City	District
TW62GD	358701	{55BDCAE6-C	330000000	20/10/2016	TERMINAL 5	SOFITEL	WENTWORTI	LONDON HE/	HOUNSLOW	HILLINGDON
NW17QX	3586981	{55BDCAE6-C	252650000	29/12/2016	265	NA	HAMPSTEAD	NA	LONDON	CAMDEN
NW17QX	358698	{55BDCAE6-C	252650000	29/12/2016	265	NA	HAMPSTEAD	NA	LONDON	CAMDEN
SE19PZ	370128	{55BDCAE6-E	140000000	08/07/2016	76	NA	UPPER GROU	NA	LONDON	LAMBETH
WC1V6JS	941660	{3914047A-8	96840522	15/06/2016	THE EYE, 100 - 110	NA	HIGH HOLBO	NA	LONDON	CAMDEN
WC1V6JS	9416601	{3914047A-8	96840522	15/06/2016	THE EYE, 100 - 110	NA	HIGH HOLBO	NA	LONDON	CAMDEN
NW12PN	2447201	{453D27A3-E	96350000	12/12/2016	THE SOLS ARMS PUBLIC HOUSE, 65	NA	HAMPSTEAD	NA	LONDON	CAMDEN
NW12PN	244720	{453D27A3-E	96350000	12/12/2016	THE SOLS ARMS PUBLIC HOUSE, 65	NA	HAMPSTEAD	NA	LONDON	CAMDEN
EC3N4SG	95887	{404A5AF4-6	84484999	29/09/2016	S G HOUSE, 41	NA	TOWER HILL	NA	LONDON	CITY OF LONDO
EC2V6BJ	91922	{404A5AF4-6	83129681	30/06/2016	139 - 140	NA	CHEAPSIDE	NA	LONDON	CITY OF LONDO
TW59NR	740720	{404A5AF4-4	79500000	16/09/2016	UNIT 20	NA	AIR LINKS INI	NA	HOUNSLOW	HOUNSLOW
TW59NS	89932	{404A5AF4-6	79500000	16/09/2016	CENTRE HOUSE	NA	VICTORY WA	NA	HOUNSLOW	HOUNSLOW
TW59NW	89938	{404A5AF4-6	79500000	16/09/2016	UNIT 2-5	NA	SPITFIRE EST.	NA	HOUNSLOW	HOUNSLOW
EC3M3BE	81708	{42A5A70A-2	78972407	24/11/2016	06-Aug	NA	FENCHURCH	NA	LONDON	CITY OF LONDO
W38SX	92063	{42A5A70A-0	77300000	27/05/2016	BLACKBURN COURT	1	BOLLO LANE	NA	LONDON	EALING
W1T7QX	514196	{4C4EE000-4	73300000	01/12/2016	247	NA	TOTTENHAM	NA	LONDON	CAMDEN
W1T7QX	5141961	{4C4EE000-4	73300000	01/12/2016	247	NA	TOTTENHAM	NA	LONDON	CAMDEN
IG12ZG	425061	{2FD36066-5	70900000	05/02/2016	03-May	NA	WINSTON W	NA	ILFORD	REDBRIDGE
NA	358700	{55BDCAE6-C	69857000	22/11/2016	THOMAS HARDY HOUSE	CIVIC FACILIT	CECIL ROAD	NA	ENFIELD	ENFIELD

Analysis

Figure 2: A group of maps representing the spatial distribution of calculated True Mean house prices in each London borough. The legend shows the colour distinction between the upper, middle, and lower third of results obtained from the True Mean price per borough (PPB)(£). Created in ArcMap.

Figure 2a: Map giving the overall distribution of True Mean price per borough.

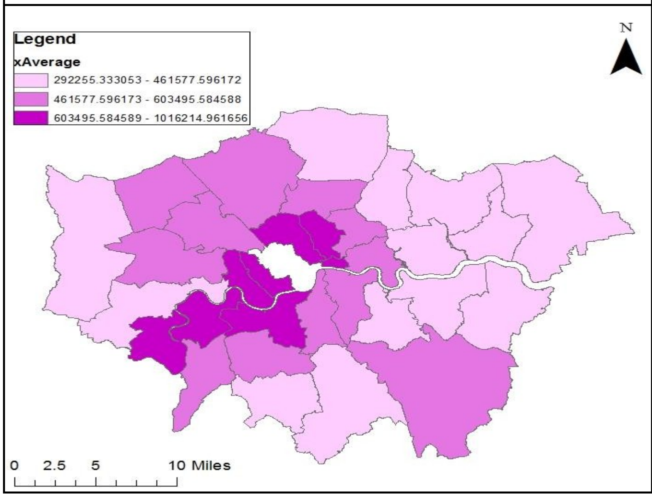


Figure 2b: Map shows the boroughs in the lowest third of PPB results.

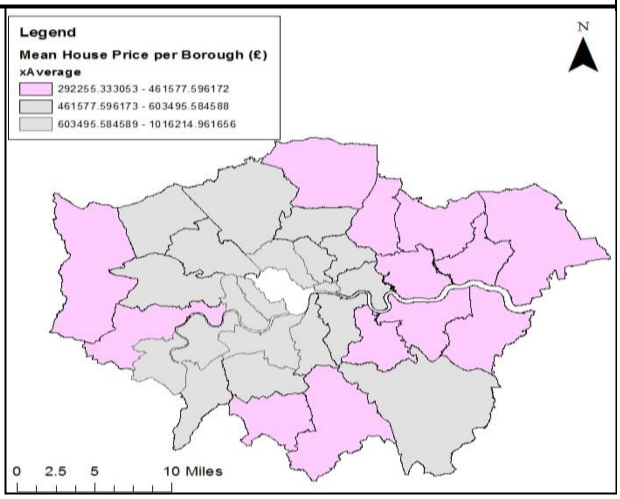


Figure 2c: Map shows the boroughs with the middle third of PPB results.

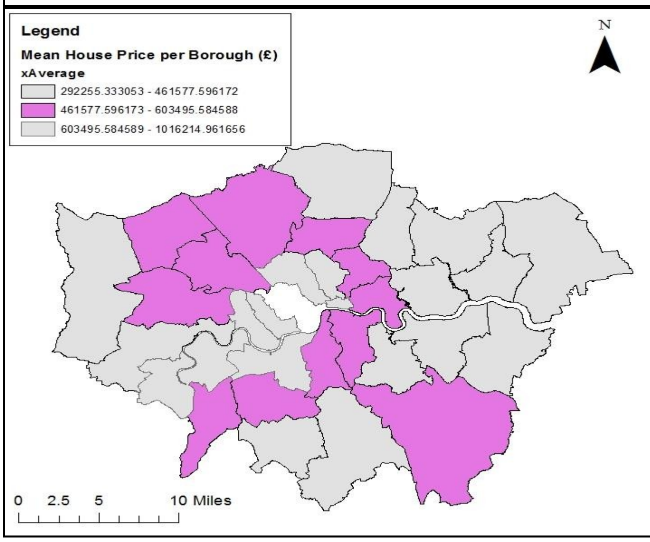


Figure 2d: Map shows the borough in the highest third of PPB results.

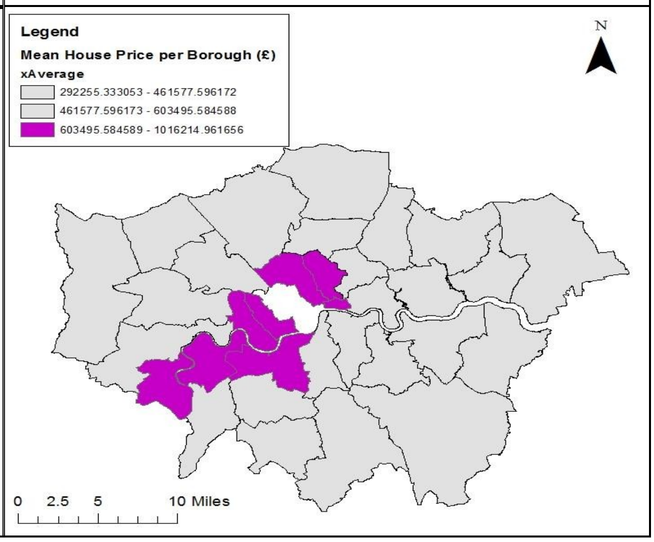


Figure 2 shows comparative illustrations of the results of the True Mean house price within each borough calculated within ArcMap. The maps within Figure 2, appear to display two significant trends. The most significant of which is the clear correlation between proximity to the center of London and mean house price, best illustrated in figures 2b and 2d, where the lowest third of mean house prices is on the London periphery and the highest third of mean house prices within boroughs located more centrally respectively. Secondly, evidence from figures 2b and 2d, suggest that in addition to the True Mean house price's decreasing in relation to distance from central London, mean house values are also higher in the West, and lower within the East of the London constituency.

Figure 3: Group of maps displaying the results from the Ordinary Kriging interpolation method performed in ArcMap.

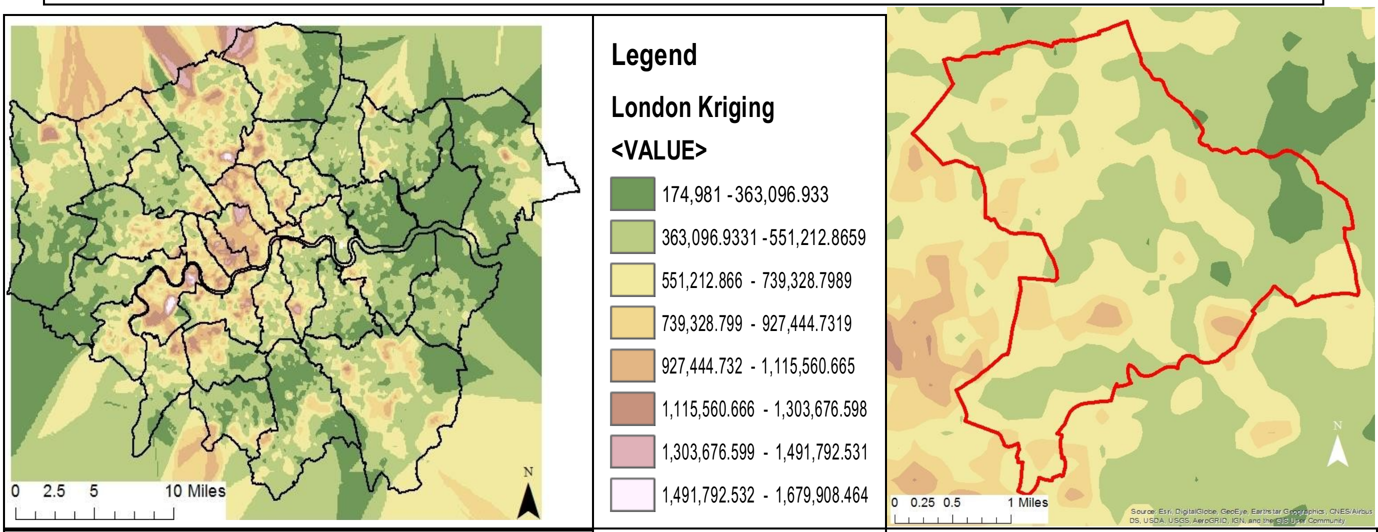


Figure 3a: Ordinary Kriging result for the entire London house price dataset within every borough, displayed at the entire London extent.

Figure 3b: Ordinary Kriging result for the entire London house price dataset, displayed at the Hackney borough extent.

OK and IDW interpolation techniques performed on ArcMap provided much greater detailed analysis of the house price dataset's distribution throughout London. Figures 3a and 3b show the OK results for the entire London dataset, at both the London and Hackney extent. The entire London dataset OK and IDW (Figures 3a and 4a) correspond with the general trend represented in Figure 2, as lower house price values (represented in green) are located further towards the London boundary extent, and adversely higher house price values can be seen centrally and towards the south-west region of the maps. In addition to this, the London IDW (figure 4a) contains a higher differential range between highest and lowest estimated house price values, yet when visualized on the map it appears to have a more smoothed affect than the London OK (figure 3a).

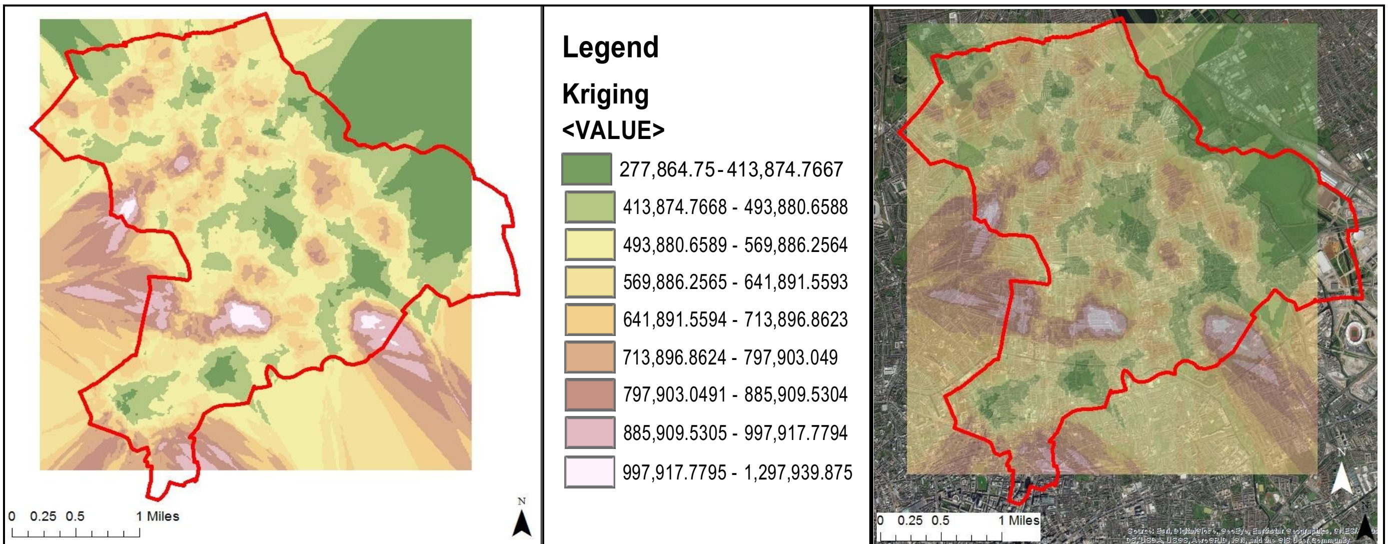


Figure 3c: Ordinary Kriging result of the house price point data within Hackney borough.

Figure 3d: Ordinary Kriging result of the house price point data within Hackney borough overlay the World Imagery basemap from ArcMap to show the result in relation to the real-world borough imagery.

Figure 4: Maps displaying the results from the Inverse Distance Weighting (IDW) interpolation method performed in ArcMap.

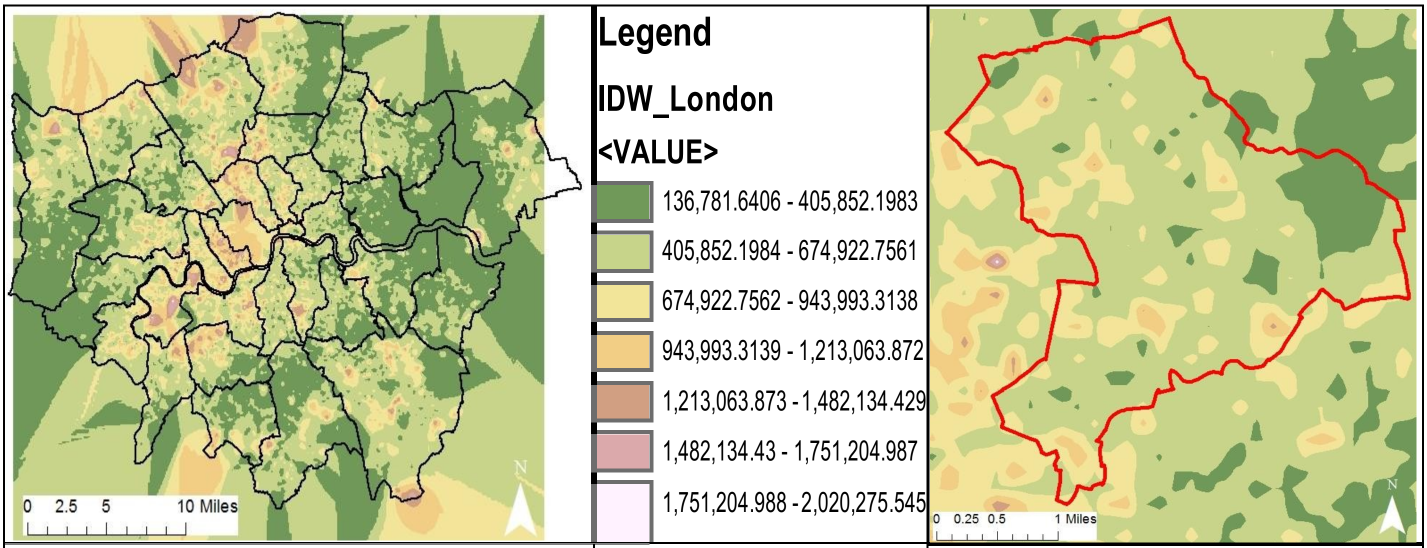


Figure 4a: IDW result for the entire London house price dataset within every borough, displayed at the entire London extent.

Figure 4b: Ordinary Kriging result for the entire London house price dataset, displayed at the Hackney borough extent.

Reasoning for studying the dataset at two contrasting extents can be observed when comparing figures 3b & c, and figures 4b & c, respectively. The comparison between these figures show the London OK and IDW, when zoomed to Hackney, are smoother and provide significantly less detail and accuracy in relation to the Hackney dataset OK and IDW results, which provide a far greater fluctuation and therefore detail of house price distribution within this borough. Further analysis can be made when comparing figures 3c and 4c, with their respective figures 3d and 4d, both coupled with figure 1b. It can be seen that the low values from both the Hackney OK and IDW interpolation methods are generally attributed to areas of land which do not contain houses situated upon. The clearest example of this is in the East of the Hackney borough, where low values (green) and even a blank area with no values correspond with a large field in the 'true-imagery' maps. It can be inferred that this is representative of the entire dataset and therefore can be assumed that the lower values results for all maps produced are largely attributed to the extent of non-urban areas such as fields.

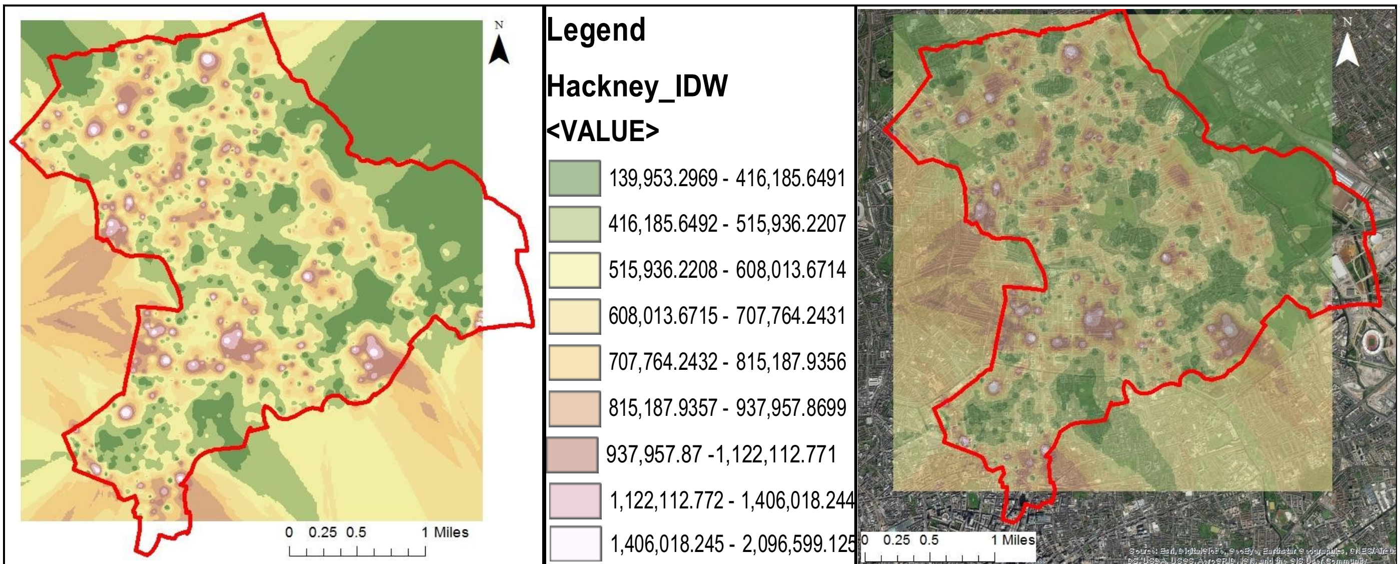


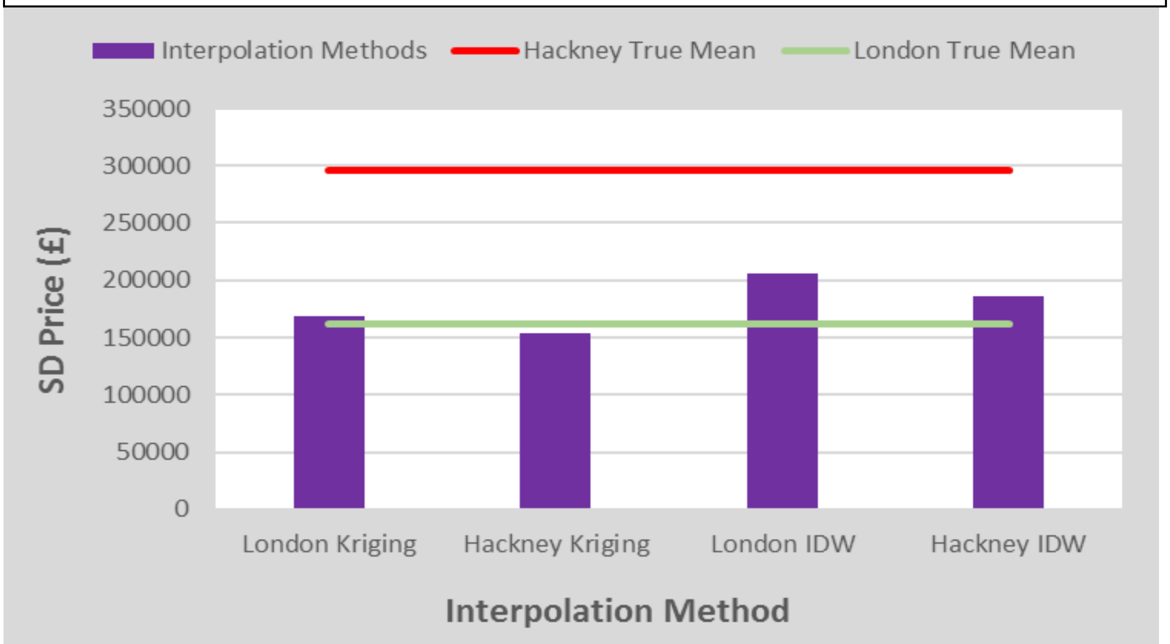
Figure 4c: Ordinary Kriging result of the house price point data within Hackney borough.

Figure 4d: IDW result of the house price point data within Hackney borough overlay the World Imagery basemap from ArcMap to show the result in relation to the real-world borough imagery.

Figure 5: Chart representing the variation in means in relation to the calculated True Means.

Figures 5 and 6 show the statistical representation of the results acquired. In figure 5 it can be seen that the True Mean for both the London and Hackney study extents provided higher mean values than both the interpolation methods. Figure 6 illustrates the standard variation of the spatial analyst methods used. The Hackney interpolation techniques yield far more representative SD than the two London interpolated results. Additionally, it can be suggested that the standard deviation has an inverted correlation with the mean, shown when a high mean value for interpolation technique is observed, a corresponding low SD value is obtained.

Figure 6: Chart representing the calculated Standard deviation in relation to the True Means values.



Discussion

The focal findings from this study emphasized the need for analysis of data to be made at both a large (whole of London) extent, and a more detailed (single borough) extent. In addition, this research indicates the limitations of the True Mean's abilities to provide detailed geospatial measurements. The True Mean model is suitable for simple overview on a large scale but doesn't provide detailed or accurate analysis. In comparison, the two interpolation techniques provided a far greater accuracy and detail than the smoothed True Mean models when representing the house prices spatial distribution throughout both London and Hackney, consequently allowing more thorough investigation to possible causes of house price distribution within an area. The True Mean averaged the entire specified dataset for that area, not accounting for variations in point values throughout the distribution. As a result, when applied to the house price data, the true mean is unable to represent how house prices and local market trends may vary over space (Cellmer, 2014; Bitter, Mulligan, and Dall'erba, 2007). This was particularly evident when comparing the different spatial analytic models to the World Imagery basemap, as it could be seen that the interpolation models provided a far better representative illustration, accounting for variations in land surface.

To conclude, the True Mean model is suitable for very simple spatial analysis, however, is very limited when conducting any detailed GIS research. Instead the interpolation models such as IDW and OK presented in this study are far more appropriate for analysis of the spatial distribution of house price data. Further study could be made as to the influence of non-urban areas on the distribution of house, and consequently, how this affects different interpolation methods.

References

Ahsan, A. and Parvez, S. (2014). Geo-statistical Methods For Spatial Interpolation in GIS. In: *International Conference on Space (ICS-2014) Organized by SUPARCO, IST, and ISNET*. Lahore: Dept of Space Science, University of the Punjab.

Bhunia, G., Shit, P. and Maiti, R. (2016). Comparison of GIS-based interpolation methods for spatial distribution of soil organic carbon (SOC). *Journal of the Saudi Society of Agricultural Sciences*.

Bitter, C., Mulligan, G. and Dall'erba, S. (2006). Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems*, 9(1), pp.7-27.

Butkiewicz, T., Dou, W., Wartell, Z., Ribarsky, W. and Chang, R. (2008). Multi-Focused Geospatial Analysis Using Probes. *IEEE Transactions on Visualization and Computer Graphics*, 14(6), pp.1165-1172.

Cellmer, R. (2014). The Possibilities and Limitations of Geostatistical Methods in Real Estate Market Analyses. *Real Estate Management and Valuation*, 22(3).

Fei, C. and Daosheng, D. (2004). Application of integration of spatial statistical analysis with GIS to regional economic analysis. *Geo-spatial Information Science*, 7(4), pp.262-267.

Goodchild, M. (1986). *Spatial Autocorrelation*. Published by Geo Books, Norwich.

Luo, W., Taylor, M. and Parker, S. (2008). A comparison of spatial interpolation methods to estimate continuous wind speed surfaces using irregularly distributed data from England and Wales. International Journal of Climatology, 28(7), pp.947-959.
Sarma, D. (2009). Geostatistics with Applications in Earth Sciences. *Springer*.

Schloeder, C., Zimmerman, N. and Jacobs, M. (2001). Comparison of Methods for Interpolating Soil Properties Using Limited Data. *Soil Science Society of America Journal*, 65(2), p.470.

Statch, A. (2009). GIS - platforma integracyjna geografii, Bogucki Wydawnictwo Naukowe, Poznań. *(Analysis and modelling of spatial structure)*, Analiza i modelowanie struktury przestrzennej.

Wong, D., Yuan, L. and Perlin, S. (2004). Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Analysis and Environmental Epidemiology*, 14(5), pp.404-415.

Xie, Y., Chen, T., Lei, M., Yang, J., Guo, Q., Song, B. and Zhou, X. (2011). Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: Accuracy and uncertainty analysis. *Chemosphere*, 82(3), pp.468-476.

Xu, Z., Guan, J. and Zhou, J. (2015). A Distributed Inverse Distance Weighted Interpolation Algorithm Based on the Cloud Computing Platform of Hadoop and Its

Implementation. 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).
Yao, J. and Stewart Fotheringham, A. (2015). Local Spatiotemporal Modeling of House Prices: A Mixed Model Approach. *The Professional Geographer*, 68(2), pp.189-201.

Yao, X., Fu, B., Lü, Y., Sun, F., Wang, S. and Liu, M. (2013). Comparison of Four Spatial Interpolation Methods for Estimating Soil Moisture in a Complex Terrain Catchment. *PLoS ONE*, 8(1), p.e54660.

Appendix

R-Code Script

#If any functions are present then look to load from library(), and if not present there then must install respective package using "install. Packages"

```
#Read in and check data for analysis------------------------------------------------
London_House_Prices_2016 <- read_csv("~london_house_prices_pp_2016.txt")
London_Postcode <- read_csv("~london_postcode_bng_lookup.txt")

#check data
View(London_Postcode)
View(London_House_Prices_2016)

#Clean the data ------------------------------------------------------
#strip out spaces
London_House_Prices_2016$Postcode <- gsub(" ", "",
London_House_Prices_2016$Postcode)
London_Postcode$Postcode <- gsub(" ", "", London_Postcode$Postcode)

#check new stripped data
View(London_House_Prices_2016)
View(London_Postcode)


#remove outliers using upper and lower quartiles---------------------------------------
#identifying quartiles
quantile(London_House_Prices_2016$Price, seq(0, 1, 0.025))
lowerq = quantile(London_House_Prices_2016$Price, seq(0,1,0.025))[2]
upperq = quantile(London_House_Prices_2016$Price, seq(0,1,0.025))[40]

#removing outliers
London_House_Prices_2016_2 = sqldf('select * from London_House_Prices_2016
where Price >= 128372')
London_House_Prices_2016_3 = sqldf('select * from London_House_Prices_2016_2
where Price <= 2123075')
```

```
#join postcode data to price paid and check join  worked---------------------------------------
#Merge
London_House_Prices_2016 <- merge(London_House_Prices_2016_3,
London_Postcode, by.x="Postcode", by.y="Postcode")
View(London_House_Prices_2016)


#read in London poly shapefile "London-house-prices-ppd-2017" and set coordinate
system---------------------------------
#read shapefile
London_poly <- readOGR(".", "London_Borough_Excluding_MHW")
plot(London_poly)
View(London_poly)


#changing London_poly Borough names to match Borough names in
London_House_Prices_2016 Borough names to enable join
London_poly$NAME <- toupper(London_poly$NAME)


#setup variables for british national gird
bng <- "+init=epsg:27700"


#Compare the "NAME" column in London_poly to "District" column in
London_House_Prices_2016
London_poly$NAME %in% London_House_Prices_2016$District


#Return rows which do not match
London_poly$NAME[!London_poly$NAME %in% London_House_Prices_2016$District]
#"WESTMINSTER" is returned, stating it doesnt have correspondent values


#rename the 'District' heading to 'NAME' to match heading in London_Poly
names(London_House_Prices_2016)[14] <- "NAME"


#Export necessary data for ArcGIS---------------------------------
#Write london house prices to CSV table to export to arcmap
write.table(London_House_Prices_2016, file = "LDN_House_Price2.csv", sep = ",",
row.names = F)
#write out london polygon to shapefile to export to ArcGIS
writeOGR(London_poly, ".", "LDN_poly", driver = "ESRI Shapefile")
```