Approximate Nearest Neighbour Techniques for Large Spatial Datasets

Martin Charlton[*], Chris Brunsdon[†]

National Centre for Geocomputation, Maynooth University, Ireland

January 10, 2017

Summary

This paper reports on the application of approximate nearest neighbour techniques to increase productivity in spatial analysis. We demonstrate the use of multidimensional binary search trees, sometimes known as k-d trees, in (i) determining spatial variations in building proximity in Ireland and (ii) two geographically weighted regression applications. The Irish data has 1.9 million records, and the GWR datasets 229 thousand and 77 thousand records. The analyses were undertaken using the R system on low cost laptops.

KEYWORDS: approximate nearest neighbour search, geographically weighted regression, database, spatial analysis

1. Introduction

Different characterisations of Big Data range from data of enormous volume, variety, veracity and velocity to that which is too large to fit into the memory of your laptop. Spatial operations on data that will fit into the memory of a laptop can provide challenges, particularly if we are using open source software. The problem is however, not new. Forty years ago a large mainframe might have 2Mb of main memory, perhaps with an external high speed disk drive which provided paged external 'virtual' memory. Bentley's (1975) paper on multidimensional binary search trees appeared at the same time, with Robinson's (1981) extension which linked the k-d-tree with the B-tree to create the k-d-b-tree. Openshaw's requirement for spatial searches in his Geographical Analysis Machine (Openshaw *et al*, 1987) and national nuclear attack simulations (Openshaw *et al*, 1983) are based on Bentley's developments.

This paper reports on spatial data manipulation and modelling with three moderately large datasets: (i) GeoDirectory – the address database for Eire, with some 2m records (ii) GB census data for Output Areas with 232000 records, and (iii) a mortgage dataset for England and Wales with 77000 records. The work was carried out using the R statistical computing and graphics environment.

2. Building spacing

Census data are~~is~~ usually aggregated to spatial units for reporting. In the UK usually resident population and household counts are available at Output Area level, and the Republic of Ireland, similar data is available at Small Area level. This does allow computation of population and household densities if the areas of the underlying spatial units are available. Ireland national address database, GeoDirectory[‡], contains a table of geocoded locations for the individual buildings ~~which~~

[*] martin.charlton@nuim.ie

[†] crhis.brunsdon@nuim.ie

[‡] www.geodirectory.ie

that have one or more associated postal addresses. This allows an alternative approach to the computation of the spatial variation in building density based on building proximity.

GeoDirectory is available in a number of different formats including Microsoft Access –a relational database management system. The tables include BUILDINGS, each entry of which provides a geocode (in Irish National Grid, Irish Transverse Mercator, and WGS84) as well a~~–~~s counts of residential and commercial delivery points. Whilst proximity to the nearest adjacent building can be computed using ArcGIS, this is more of a challenge in R; there are 1.9m records in the table. A brute force approach would be to compute the distances between every pair of buildings – this would be unfeasibly slow. There are no obvious ways in which data in two dimensions can be sorted to allow proximity to the determined. ~~The~~ One solution is to use a k-dtree, and Arya and Mount's (1993) ANN library provides a convenient implementation for a 2-d tree.

Tables from an Access database can be read into R using SQL queries available through the *RODBC* package, and the RANN package provides a function to build and search a 2-d tree. For the distance to the nearest adjacent building we require the 2-nearest neighbours (2NNs): the first NN is the building itself, and the second is the nearest adjacent building. The function returns not only the row numbers of the NNs but also their distances.

The results are obtained quickly – on a low-cost laptop with a 2.2Gz processor finding the 2NNs for 2m records takes about 8 seconds, about 13 seconds for 3m records, about 18 seconds for 4m records, and about 25 seconds for 5m records, and 120 seconds for 20m records. The line in Figure 1 suggests the relationship is linear.

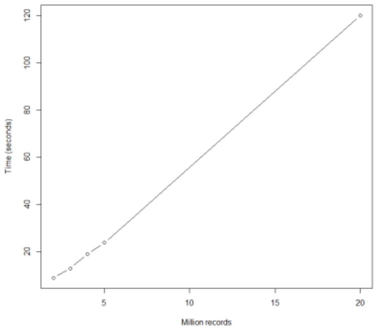

Figure 1 Search times for various dataset sizes.

The timings in Figure 1 are based on random data. Experiments with clustered data suggested they also line on the same line.

Reading the 1927810 records from the GeoDirectory BUILDINGS table takes about 10 seconds, and the nearest neighbour search takes about 5 seconds. The median building spacing is 10.0m, the mean is 36.0m and the maximum is 3.82km! There is, however, considerable spatial variability from a median of 6.0m in Dublin City to 43.2m in County Leitrim. Summarising the medians at Electoral Division (ED) level (n=3409)
gives the result in Figure 2. The distinction between the main urban areas (Dublin, and the cities of Galway, Limerick, Cork and Waterford) and their surrounding regions are very clear, and the local rural maxima that represent the network of rural towns around Ireland.

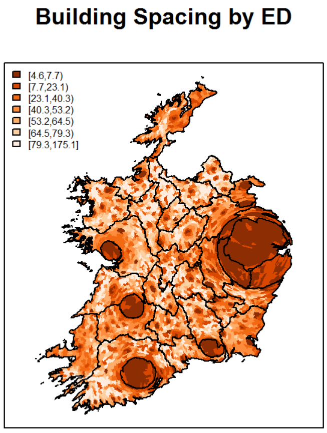

Figure 2 Median spacing by Electoral Division

3. Spatial modelling

A second example concerns the scalability of *geographically weighted* techniques. Software made available following the publication of Brunsdon et al (1996) uses a brute force approach to compute the geographical weights. This limits the applicability of the geographically weighted techniques to relatively small datasets. The ArcGIS Geographically Weighted Regression tool will handle datasets with several thousand observations but is limited to regression models with a Gaussian error term. Other implementations in R also use brute force methods – usually requiring the re-computation of the distances in each iteration of the estimation, which can be wasteful in terms of computing resources. However, 2-d trees offer a solution for dealing with moderately large datasets efficiently.

The issue arises in the computation of the geographical weighting. If a Gaussian kernel is used the distances between every pair of observations are required; if Euclidean distances are used, then for n observations $\frac{1}{2}n(n-1)$ distances have to be computed for a complete estimation of the model. During the calibration phase, these distances must be available for each iteration. Pre-computing and storing the distance matrix is possible for smaller problems, but the amount of disk space required is O(n^2); ~~the~~ limitations on RAM impose more severe boundaries.

When an adaptive kernel is used, the *h* nearest neighbours are required, together with their distances, *h* being the bandwidth. The brute force approach computes the distances from the current observation to all the others, sorts then in order, and then extracts the distance for the 2nd to *h*th nearest neighbours, which are then used to compute the weights. Recall that the ANN library provides these distances as ~~a result~~an output of the k-d-tree search, which yields two advantages. The first is the computation of the distances for the *h*NNs is far faster, and the second is that much larger problems can be accommodated as a by-product.

3.1 Modelling educational attainment in the UK

The dataset with socio-economic variables relating to educational attainment in the counties of the state of Georgia in the USA is well-known. A similar dataset was put together for the 232296 Output Areas in the UK with four variables:

1. Proportion of residents educated to Level 4 or higher

2. Proportion of economically active who are unemployed
3. Proportion of households who are social renting
4. Proportion of economically active using public transport for journey to work

Educational attainment was modelled as a function of the other three. The calibration required 13 iterations, to produce a bandwidth of 66 OAs. During the calibration process, few diagnostic statistics are computed, and the hat matrix is not available, so the calibration criterion is to minimise a *leave-one-out* cross-validation statistic. Estimation of the complete model took 4 minutes and 20 seconds.

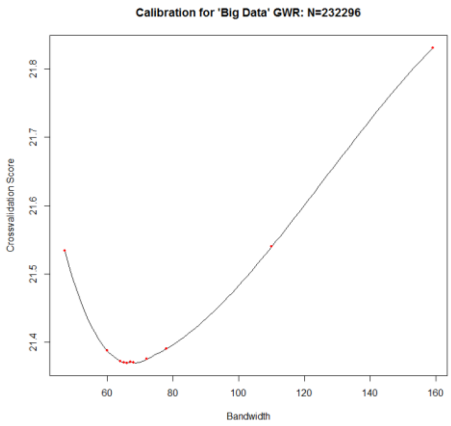

Figure 3 Calibration

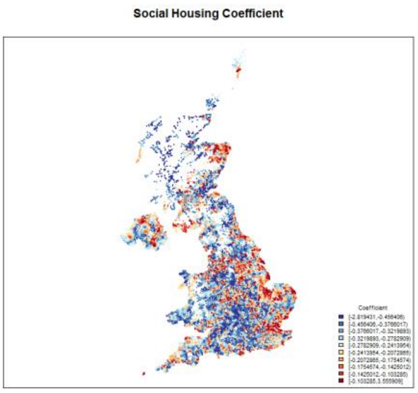

Figure 4 Social Housing Coefficient

Figures 3 and 4 show the shape of the calibration function – the process is based on a Golden Section search, and converges relatively quickly. The Social Housing Coefficient is mostly negative – research is needed on the most appropriate means of dealing with the issue of multiple testing – the output implies 929184 separate significance tests of which we would expect 46459 to be significant at random if the 0.05 significance level is chosen and there was no association.

3.2 Modelling residential property prices

The final example is based on an analysis the authors ~~first~~ originally conducted in 1999. The dataset consists of 77299 anonymised mortgage records for property sales in 1990 in England Wales. We constructed a hedonic model to predict the spatial variation in the property price. The predictor variables include floorspace, building type and age, presence of a double garage and the proportion of residents in professional and managerial occupations in the neighbourhood. The type, age and garage variables are in the form of dummies (0/1), and the omitted classes are interwar bungalows.

Collinearity has been identified as an issue in GWR, so we compute the local condition number for each estimation. The calibration took 1550 seconds, and the estimation of the model and diagnostics required an additional 280 seconds, just over half and hour in total.. This compares with the 120 hours it took to calibrate the model on a dedicated Sun server in 1999. Note that the time to run the model is not only a function of the number of observations, but is also related to the number of predictor variables – the estimation requires the inverse of a *p* by *p* matrix to be computed for each observation.

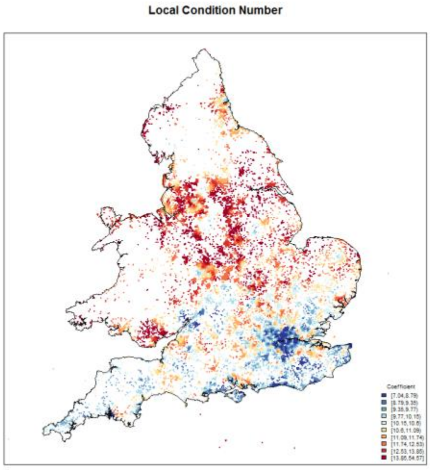

Figure 5 Local Condition Number

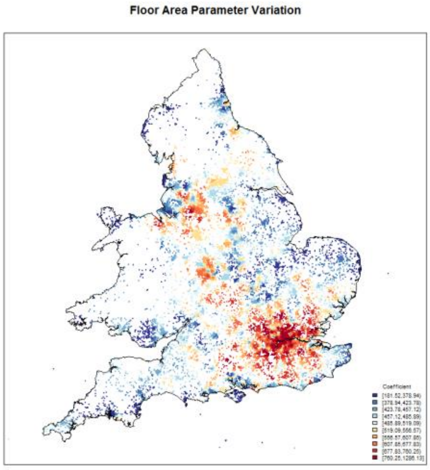

Figure 6 Floorspace Parameter Variation

The spatial variation in the local condition number suggests there are slightly higher levels of collinearity outside London, the south-east and south-west. However, only 1% of the CNs are above 18.9, and 50 (0.065%) are greater than 30, so collinearity is not really an issue with these data. The floorspace parameter represents the price per square metre of floorspace. The primacy of London is very noticeable, and there are higher prices in the core area north west of London towards Manchester. In contrast with Manchester, prices are low in Liverpool.

4. Conclusions

The use of spatial indexing in 'large-ish' data analysis would appear to be beneficial. We can conduct analyses that might be unfeasible with brute force methods; and techniques such as GWR appear to be linearly scalable.

5. Acknowledgements

GeoDirectory is made available through An Post Ireland.

6. Biography

Martin Charlton is Senior Research Associate at the National Centre for Geocomputation, Maynooth University Ireland.

Chris Brunsdon is Professor of Geocomputation at the National Centre for Geocomputation, Maynooth University Ireland.

References

Arya S and Mount D M (1993), Approximate nearest neighbor searching, *Proceedings of the 4th Annual ACM/SIAM Symposium on Discrete Algorithms (SODA'93)*, 271-280.

Bentley J L (1975), Multidimensional binary search trees used for associative search, *Communications of the Association for Computing Machinery*, 18, 309-517.

Openshaw S, Charlton M, Wymer C and Craft A W (1987), A mark I geographical analysis machine for the automated analysis of point data sets, *International. Journal of Geographical Information Systems*, 1, 335-358

Openshaw S, Steadman P and Greene O (1983) *Doomsday: Britain after nuclear attack*, Oxford: Basil Blackwell

Robinson J T (1981) *The k-d-b-tree: a search structure for large multidimensional dynamic indexes*, Research Report CMU-CS-81-106, School of Computer Science, Carnie-Mellon University