A moan, a discursion into the visualisation of very large spatial data and some rubrics for identifying big questions

A. Comber[1], C. F. Brunsdon[2], M. Charlton, R. Harris[3]

[1] School of Geography, University of Leeds, LS2 9JT, UK
Email: a.comber@leeds.ac.uk

[2] NUI Maynooth, Maynooth, Co Kildare, Ireland
Email: {christopher.brunsdon; martin.charlton} @nuim.ie

[3] School of Geographical Sciences, University of Bristol, BS8 1SS.
Email: rich.harris@bris.ac.uk

Abstract

This short paper links 2 areas of big data science in the context of GIScience: inferential analysis and visualisation. It discusses ideas around integration and analysis of large, spatial referenced datasets and considers how results of these can best be visualised. It advocates a critical approach to big data visualization and warns of the inherent dangers of simply identifying patterns, whether through data mining, modeling or visualization. It adds to ongoing debates by suggesting techniques and rubrics, possibly even hinting at a manifesto.

1. Introduction

There is an increasing amount of data of all kinds available to scientists that provide opportunities to gain novel insights about all kinds of phenomena. The availability of these data is being driven by 2 factors. 1) The large amount of open data and wider recognition of the value that can be added to that data (Molloy, 2011) by linking it to other data and by developing novel data analyses; 2) The many new forms of data generated every day by citizens, either passively or actively (See *et al.*, 2016) on GPS- and web-enabled tablets, devices has resulted in an explosion of citizen contributed, crowdsourced or volunteered data.

Much has been written about the characteristics of these very large datasets: from the 3 or 5 or is it 7Vs? to the 3Ds (Dynamic, Diverse, Dense to which Dirty should be added), and perhaps more interestingly, their existence has stimulated a number of theoretical and practical considerations. This ranges from the need to revisit classic measures of and tools for statistical inference (Brunsdon, in press) to the need to redesign some of the more commonly used software tools to handle the data volumes. The role of GIScience relates to location, which may be precise in the form of latitude and longitude or approximate for example using a small census area reference or a post-code. However, despite being in this age of so called 'Big Data' the real challenge is to identify and answer 'Big Questions' which so far the research community, including the GIScience community, has failed to do.

2. Large quantities of spatially referenced data

There are large quantities of spatially referenced data of many different types, describing many different phenomena. This provides opportunities for new forms of knowledge. A typical data-mining / computer science approach is encapsulated by the following quote: '*Scouring databases and other data stores for insight is often compared to the proverbial search for a needle in a haystack, but … big data turns that idea on its head*' and quoting Viktor Mayer-Schönberger '*With big data, we don't know what the needle is. We can let the data speak and use it to generate really intriguing questions*'[1].GIScience offers suites of

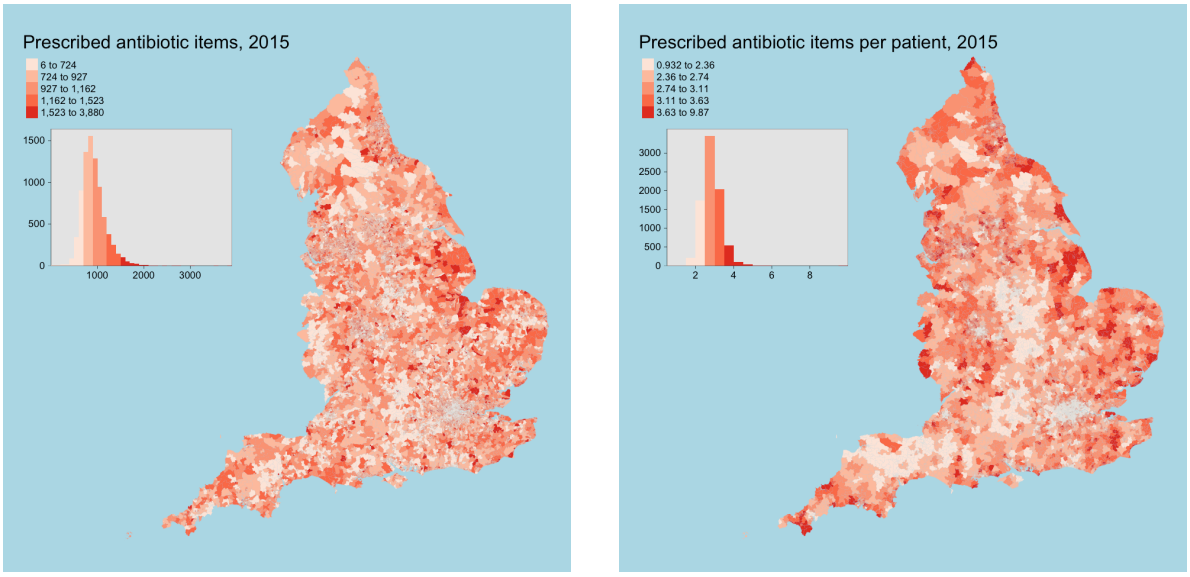
[1] http://data-informed.com/big-datas-value-much-larger-than-specific-business-questions/

methods and techniques for integrating such data over defined geographic areas (van der Zee and Scholten, 2014), for example by summing the counts of some phenomenon over a census areas and for analysis. Consider for example, the medical prescription data provided by the UK government for England (Figure 1). It records individual prescriptions, the prescribing GP practice or hospital, the cost and month of prescribing, with 10.1m records for January 2015. The GP practice postcode is provided separately and can be used to locate the data.

```
>
> head(data.1)
  SHA PCT PRACTICE        BNF_CODE                                BNF_NAME ITEMS   NIC ACT_COST QUANTITY PERIOD
1 Q44 RXA    N81646 0102000N0AAABAB Hyoscine Butylbrom_Tab 10mg              1  1.13     1.16       21 201501
2 Q44 RXA    N81646 0401010Z0AAAAAA Zopiclone_Tab 7.5mg                     15  3.28     4.72       57 201501
3 Q44 RXA    N81646 0401020K0AAAHAH Diazepam_Tab 2mg                        35 99.91   104.64     2662 201501
4 Q44 RXA    N81646 040201060AAALAL Olanzapine_Tab 15mg                      3  1.29     1.53       21 201501
5 Q44 RXA    N81646 0403010B0AAAHAH Amitriptyline HCl_Tab 25mg              16  1.52     3.20       38 201501
6 Q44 RXA    N81646 0403010B0AAAIAI Amitriptyline HCl_Tab 50mg              44  4.04     8.68      101 201501
>
```

Figure 1. Example of prescription level data.

Links to other information are needed to extract added value. Consider the aim identify factors related to Antibiotic resistance (ABR) from prescriptions. One approach is to using data mining to identify possible relationships between antibiotic prescribing patterns with a plethora of socio-economic factors. Another is to consider specific factors, for example, related to ill-health, deprivation, old age etc, on the basis that these social groups may have higher anti-biotic prescribing rates and therefore may be more likely to exhibit ABR at some point in the future. To illustrate this, prescribing data for 2015 was linked to census data for the 32,844 lower super output areas (LSOAs) in England plus 7 random created variables (r1 to r7). Figure 2 shows antibiotic prescribing rates items per patient for LSOAs in England.



a) b)

Figure 2. a) Antibiotic prescribing rates, and b) a cartogram of the same for LSOAs

When a large number of socio-economic attributes are used to construct a predictive model of prescribing rates (items per census area), many factors may be significant and with a reasonable model fit (Figure 3a). Figure 3b maps the LSOA outliers. This raises 2 problems. First, the potential meaninglessness of the variables identified as significant although in this case we could try to explain many of them. But the point is that this would be a *post hoc* rationalization. It is easier to tell convincing stories based on what is found whilst

oblivious to the potential sensitivity of the results to the methods by which the data have been compiled and collected. The traditional modelling framework doesn't work well for big data but without that framework what are we left with? How do we establish what is or isn't credible? Do we develop routines to examine specific groups of data in order cluster areas? A geo-asthenic classification on the combinations of drugs prescribed, for example? PCA could help here as potentially useful as inputs to PAM (Partitioning Around Medoids) or k-means.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.131e+01  7.376e+00   4.245 2.20e-05 ***
r1               -1.711e-04  7.882e-05  -2.171 0.029924 *
r2               -1.681e-04  7.881e-05  -2.133 0.032951 *
r3               -8.712e-05  7.882e-05  -1.105 0.269056
r4               -2.665e-05  7.882e-05  -0.338 0.735321
r5               -1.073e-05  7.882e-05  -0.136 0.891678
r6                3.541e-05  7.885e-05   0.449 0.653318
r7               -1.417e-05  7.884e-05  -0.180 0.857377
LTSickDisab      -9.079e-01  9.736e-02  -9.326  < 2e-16 ***
Stud              2.219e-01  3.137e-02   7.074 1.54e-12 ***
Inactive          9.933e-01  5.584e-02  17.788  < 2e-16 ***
InactiveOther    -1.300e+00  6.845e-02 -18.992  < 2e-16 ***
Pop               1.597e-01  1.531e-02  10.431  < 2e-16 ***
LL                5.586e-02  3.203e-02   1.744 0.081173 .
BadHealth        -7.740e-02  6.329e-02  -1.223 0.221360
EconPop          -9.556e-02  1.782e-02  -5.363 8.26e-08 ***
Unempl            1.930e-01  5.860e-02   3.293 0.000994 ***
Retired          -3.513e-01  6.239e-02  -5.631 1.80e-08 ***
StudInactive     -1.445e+00  5.774e-02 -25.024  < 2e-16 ***
IMD               5.327e+00  5.128e-01  10.389  < 2e-16 ***
Income           -5.289e+02  6.944e+01  -7.616 2.68e-14 ***
Employment        1.802e+01  6.827e+01   0.264 0.791798
Education         9.718e-01  1.095e-01   8.872  < 2e-16 ***
HealthDisability  7.094e+01  2.215e+00  32.028  < 2e-16 ***
Crime            -2.515e+01  1.674e+00 -15.019  < 2e-16 ***
BarriersHousing  -1.635e+00  1.177e-01 -13.890  < 2e-16 ***
LivingEnv        -1.588e+00  8.796e-02 -18.052  < 2e-16 ***
IncomeChildren   -1.010e+02  2.288e+01  -4.416 1.01e-05 ***
IncomeElderly    -3.261e+02  1.734e+01 -18.809  < 2e-16 ***
patients          3.755e-01  3.865e-03  97.150  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 135.4 on 32814 degrees of freedom
Multiple R-squared:  0.6592,    Adjusted R-squared:  0.6589
F-statistic:  2188 on 29 and 32814 DF,  p-value: < 2.2e-16
```

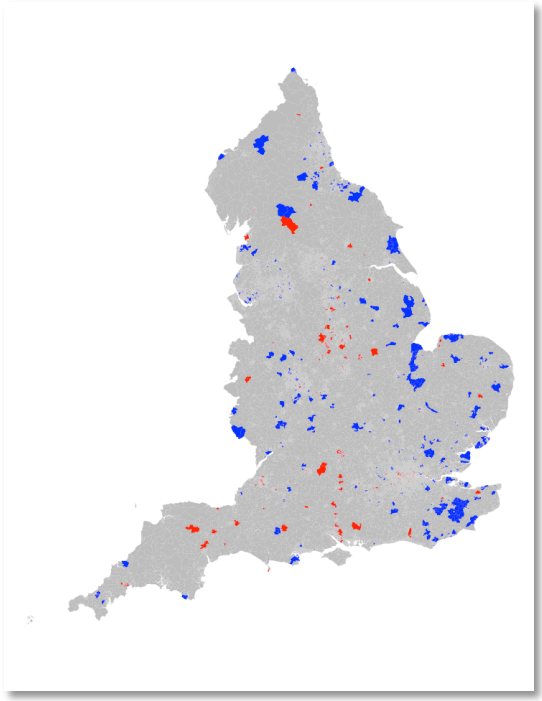


Figure 3. A model of prescribing rates (coefficient estimates) and a map of outliers

Second, the difficulty in identifying spatial patterns or trends using standard mapping approaches, so oft cited as being the panacea for big spatial data integration (Keim *et al.*, 2013). The problem is that LSOAs with larger areas dominate visually. Cartograms could be used to rescale the LSOA areas proportional to the LOSA coefficient estimate for long term sick and disabled (‘LTSickDisab’) applied to the value for each LSOA (Figure 4a). But it is difficult to interpret them because the LSOAs themselves are unfamiliar. This is in contrast to the classic cartograms of GDP developed in the 1970s and 1980s to highlight income inequalities between developed and less developed countries. So what are the options here? One might be to use regular structures such as hexbins to pool the data as in Figures 4b but to continue to undertake the analysis at the finer resolution and then to report the results using the new structures. Another approach being investigated by the authors is to transform the cartogram data but to maintain the topology of the original structures.

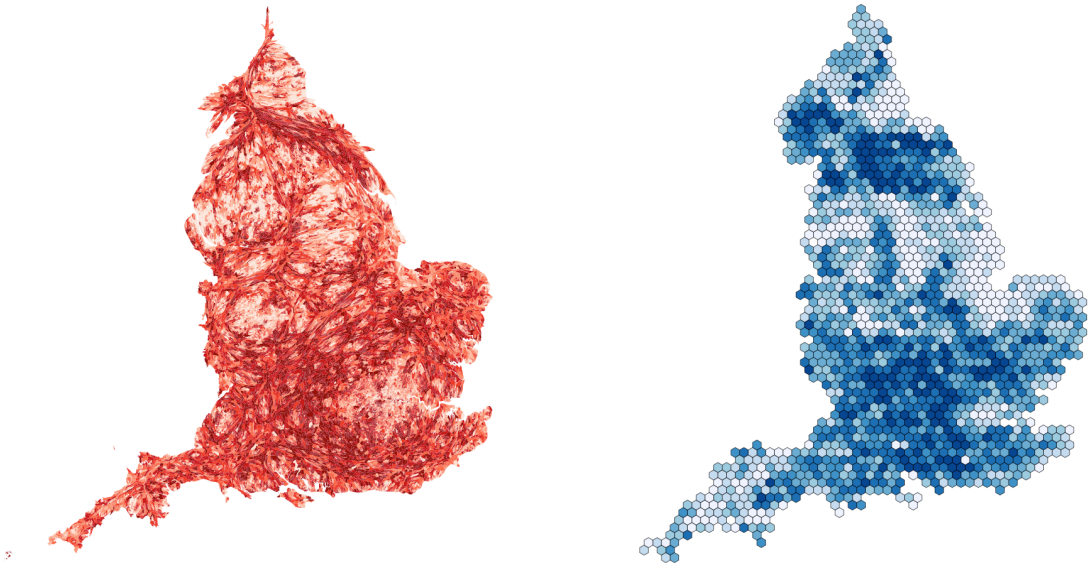
3. Rubrics

The map is a very powerful and familiar communication tool. We use them or any other visualisation to convey some information and to do that often involves also providing some critical context to the information to give salience. Context in mapping often seeks to provide a feeling of place using information that is familiar and therefore interpretable by the target audience. Maps that do this are persuasive and mapping geographical data will reveal geographical differences. However, it has long been understood in statistics that differences between values, between places are not, in themselves, evidence of anything particularly

important. Frequentist statistical testing (e.g. the t-test) puts the emphasis on statistical significance, irrelevant in the context of big data where everything will be significant due to sample size. So perhaps we need to consider how we meaningfully visualise (map) something that is meaningful. Do we for example consider dynamic linked graphics (Wood *et al.*, 2011) with immersive displays (Turkay *et al.*, 2014) to visually explore the data dimensionality?

What is needed is a balance between effective visualisation and acknowledgment that that visualisation becomes part of the problem if it legitimises dubious data and dubious research practices. This suggests the need for critical approaches big data visualisation. We need better protocols for visualisation and analysis but also to recognise that traditional visualisations may obscure detail that is important locally, although perhaps not globally, in similar way to the need for novel inferential approaches to replace those that assume small samples and situations where nil and null hypothesis testing are not relevant, such as those under big data.

What matters is whether we are learning anything particularly useful through analysis and visualisation of big data. This requires consideration of how and why the data are generated. Perhaps a test of result credibility is needed, for example by analysing random samples of the big data to see if the same results are generated. Or testing against a geographical shuffling of the data. Are similar spatial patterns observed (albeit in different places)? It is not good enough to simply take a data set, discover something and then to map and publish the results. This only serves to increase publication bias and if we only ever publish things that purport to be surprising but are actually due to luck and/or dodgy data then we will continue to publish the exception rather than the rule, which will not benefit wider society.



a) b)

Figure 4. a) shaded cartogram of showing areas where antibiotic prescribing is related to long term sick and disabled, and b) aggregated over 10km hexbins

Spatial big open data undoubtedly has the capacity to support better science and to benefit of society through increased transparency, reproducibility, efficiency etc (Molloy, 2011) and protocols to knit data together have been suggested (e.g. van der Zee and Scholten, 2014). Some have suggested a role for digital geography/GIScience not only in integrating and analysing large amounts of spatial data, but also in the engagement of the '*hard work of theory*' (Pickles, 1997, p. 370) to wrestle information control away from a very small group of corporate entities (Thatcher, 2014). However, others have noted the gaps and myths that exist between the rhetoric of big spatial data analyses and the reality (Janssen *et al.* 2012). Of

course, experimental design is also important and inferential statistics originally considered this as a part of the process of discovery, in the following way:

1. Formulate a research question.
2. Identify what data to collect and how to collect it.
3. Perform some statistical tests to determine whether any effects/associations are unlikely to have occurred by chance.
4. Get an answer to the question.

The big data paradigm turns experimental design / inferential theory on its head (the minus signs are intentional):

-1. Collect lots of data about anything.
-2. Perform some kind of data mining.
-3. Get some kind of answer.
-4. Decide what question it was an answer to.

The danger is that some researchers advocate omitting stage -4. This is problematic: if the research question is not specified then the results are answers to arbitrary questions. The means that instead of finding an answer to 'does drug X reverse type 2 diabetes' you may get an answer to a completely irrelevant question (e.g. 'do grey cats fart more often than ginger ones?') and frequently lacking an inferential dimension (information on cats farting might be spurious). We probably do need to exploit big data, but perhaps we should develop better protocols for their analysis. That is we should have some idea of what questions are important and have a better way of validating any findings. Considered, representatively robust visualisations (i.e. that communicate space and value, locally and globally) may help in this.

Until we act in this way then we cannot know whether we should expect the Big Questions to be deep in the Big Data? Or whether playing with Big Data will help us to answer Big Questions we currently have (and regardless of the answer, the MAUP does not go away). Our final observation in relation big data is that if the aim is to find a needle in a haystack then making the haystack bigger does not make the job any easier. If we don't know what kind of needle that we are looking for it helps even less. Critiques of question and context free spatial data analysis, so called *databating*, persist.

Acknowledgements

We acknowledge the stimulation of the IGU Applied Geography workshop on *the Application of Big data in Geography*, Rhodes 7-10th May 2016 that led to extensive discussions of these ideas with the bones being put together at 34,000 feet on EZ8790.

References

Brunsdon, C (in press). Quantitative methods II: Issues of inference in quantitative human geography. *Progress in Human Geography*

Janssen M, Charalabidis Y & Zuiderwijk, A (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *29*(4), 258-268

Keim, D, Qu, H, & Ma, KL (2013). Big-data visualization. *Computer Graphics &Applications*, *33*(4): 20-21

Molloy JC (2011). The open knowledge foundation: open data means better science. *PLoS Biol*, *9*(12) e1001195

Pickles J (1997). Tool or science? GIS, technoscience and the theoretical turn. *Annals of the AAG*, 87: 363-372

See L et al (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information *ISPRS International Journal of Geo-Information*, 5: 55

Thatcher J (2014). Big Data, Big Questions| Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data. *International Journal of Communication*, *8*:19

van der Zee E, & Scholten H (2014). Spatial dimensions of big data: Application of geographical concepts and spatial technology to the internet of things. In *Big Data and Internet of Things* (pp. 137-168). Springer.

Turkay C, Slingsby A, Hauser H, Wood J & Dykes J (2014). Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data. *IEEE Trans on Vis and Computer Graphics*, 20: 2033-2042

Wood J et al (2011). Visualizing the dynamics of London's bicycle-hire scheme. *Cartographica*, 46: 239-251