Crowd sourced vs centralised data for transport planning: a case study of bicycle path data in the UK

Robin Lovelace*[1]

[1]School of Geography, University of Leeds

November 13, 2014

Summary

This paper seeks to test the often mooted hypothesis that distributed, user-contributed 'crowd sourced' GIS data will eventually supercede the traditional centralised geographic data model. The empirical basis used to explore this question is a couple of national-level datasets on a specific topic: bicycle paths in the UK. Open Street Map data represents the crowd-sourced model; Ordnance Survey's Urban Paths layer represents the centralised model. To assess the quality of each dataset, an array of tests was used, from narrow tests of accuracy against aerial photography, to more subjective tests of usability and practical utility. Overall it was found that the OSM data model won on the majority of criteria. However, it must be noted that this is a niche area. If the crowd-sourced data model is to triumph in more mainstream areas it needs to ensure much greater community 'buy-in', for example through compulsory engagement with Open Street Map for educational and citizenship purposes at school.

KEYWORDS: volunteered geographic information, Open Street Map, accuracy, bicycle paths.

1 Introduction

The relative merits of different geographic datasets has long been a source of academic interest. In the sixth century BC Herodotos, the 'first map maker', criticised the inaccuracies of military maps for (Roller 2010). M. F. Goodchild and Gopal (1989) brought these critques into the 20th century, in an edited compilation of papers on accuracy in spatial data. This and other work has led to the emergence of 'testing spatial accuracy' as a sub-genre within the GIS discipline (e.g. Hunter and Goodchild 1995; Thapa and Bossler 1992; Arnold and Zandbergen 2011).

Throughout the vast majority of Geography's long history, documented geographic information has been created by only a tiny subset of the population. Millitary planners accross all cultures, Roman road builders and, more recently, cartographers armed with specialist and valuable equipment were among the 'geographic 1%' with the priveledge of filling in the blanks in humanity's perception

*r.lovelace@leeds.ac.uk

of the world. Since Open Steet Map began freely accepting data submissions from amatures in 2002, we have been living in an era of Volunteered Geographic Information (VGI) (Goodchild 2007). Growth in these new data sources far now far outstrips the incremental and accretive growth in official map source originating from trusted organisations. In fact, the British Ordnance Survey and the US Geological Survey (two of the world's most respective providers of geographic information), have already considered integrating the insights gathered from an army of 'citizen sensors' into their existing systems. The rapidly evolving landscape of VGI from Social Media (VGI-SM), which is becoming increasingly widespread with the penetration of GPS-equipped smart phones in emerging markets worldwide, only adds to the buzz surrounding crowd-sourced datasets.

Though the scope of these introductory comments are broad, the empirical aims and objectives of the study are narrow. The empirical aims of the project are as follows:

1. Describe the cycle path data in Open Street Map and test the hypothesis that it is a good source of data on cycle paths in Great Britain, suitable for academic research.
2. Compare the OSM cycle path data with a proprietary dataset.
3. Describe and explain the spatio-temporal distribution of additions to the the cycle path dataset and investigate how this corresponds with investment in cycle schemes overall.
4. Use the results to discuss the use of crowd-sourced data in mission-critical professional planning applications.

2 Data

2.1 Open Street Map data

The cycle paths were extracted from the up-to-date and compressed 'british-isles' .osm.pbf file (746 Mb), downloaded from OSM services provider Geofabrik.de. This was converted for processing into a raw .osm (14.7 Gb) text file, a variant of the xml filetype, using the command-line program **osmconvert**, accessed through the Linux command line.

To take only navigable paths, the first stage was to subset all *ways*. This led to a 4.9 Gb file containing 3.35 million features. Of course, only a small sample of this large dataset is related to cycling (Fig. x). A careful sampling strategy was therefore used to ensure all bicycle paths, according to OSM data, were captured and whilst minimising data that do not represent bicycle paths. As shown in Fig. x, the majority of the route network in OSM is *potentially* cyclable, with most residential, service and unclassified highways being suitable for bicycles. However, we are not interested in where one *can* cycle in this dataset: the focus is on where local authorities and other organisations have invested in sustainable transport by constructing bicycle paths. The 'cycleway' tag is the 11th most commonly used tag, comprising 1.6% of all highway tags in Great Britain (Figure 2.

Tagged cycleways comprise only a small proportion of recommended *cycle routes* in OSM, which are defined using a wide variety of both tags and relations. This means that extracting all cycle routes cannot be acheived with a simple one-line operation. The following command, using the

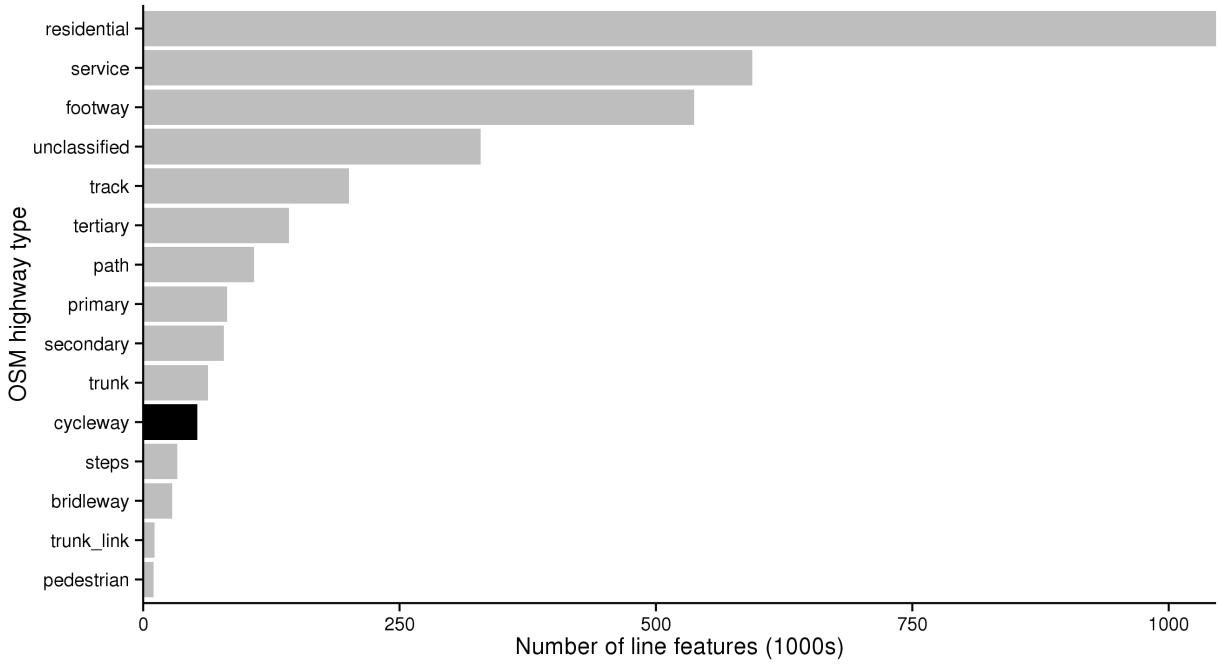

Figure 1: Number of line features by the high level 'type' tag in the UK. The only type guranteed to be a cycle path is 'cycleway', although many other 'types' can also be tagged as bicycle paths in other ways.

command-line Java program `osmfilter`, for example, will extract only a sample of the total length of cycle paths stored in Open Street Map:

```
osmfilter data.osm --keep="highway=cycleway" >bcways.osm
```

This is because the 'cycleway' attribute is only one of the possible *tags* used to represent cycle paths. There are many additional tags from a variety of variables that can but used to identify cycle paths, each with subtly different (and sometimes overlapping) meanings (Table 1).

Table 1: Overview of the most commonly used cycling-related variables and tags in Open Street Map. Note that percentages in bold refer to the proportion of the *network* (tag = * is a wildcard) while the denominator for the specific tags is the count of non-empty rows *for that variable*. Quoted text is taken from the OSM wiki.

Variable	Tag	Number	Percentage	Meaning
highway=	*	3351300	100.0	
	cycleway	52672	1.6	Dedicated bicycle paths
bicycle=	*	128493	3.8	How permitted are bicycles on the path
	yes	79586	61.9	Bicycle can ride here, not necessarily bike path
	no	24162	18.8	Bicycles are not permitted
	designated	16036	12.5	"Where a way has been specially designated"
	dismount	4015	3.1	You must push your bike by law
cycleway=	*	34085	1.0	"an inherent part of the road"
	no	16472	48.3	No bicycle path
	track	6926	20.3	Cycle path separated from the road
	lane	8831	25.9	On road/shared use bicycle path
	shared	1560	4.6	"Cyclists share space with other traffic here"
	opposite_lane	806	2.4	A 'contraflow' cycle lane
	share_busway	502	1.5	Bicycles share space with buses
	opposite	414	1.2	Bicycles can travel either way
lcn=	*	15320	0.5	Local cycle routes
	yes	13193	86.1	Completed
	proposed	1536	10.0	Proposed/under construction
ncn_ref=	*	12684	0.4	Reference of National Cycling Network paths
	1	1179	9.3	NCN route 1
	5	1003	7.9	NCN route5
ncn=	*	3313	0.1	Is part of the National Cycling Network
	yes	2368	71.5	Completed
	proposed	888	26.8	Proposed/under construction
towpath=	*	2795	0.1	The route is along a canal towpath

It is important to note that although the tags presented in figure x are *related* to cycling, they certainly do not all imply the presence of a bicycle path.

Based on a careful reading of the 'official' tag description from the OSM wiki page, and visual inspection of the paths overlaying aerial photography, a list of these 'cycle infrastructure' tags was compiled. These were:

```
"highway"  =  'cycleway' OR "bicycle" = 'designated' OR
"cycleway" = 'track' OR "cycleway" =  'lane' OR
"cycleway" = 'shared' OR "cycleway" = 'opposite_lane' OR
"cycleway" = 'opposite_track' OR "cycleway" = 'segregated' OR
"cycleway" = 'shared_lane' OR "cycleway" = 'yes' OR
"cycleway:left" = 'lane' OR "cycleway:left" = 'track' OR
"cycleway:right" = 'lane' OR  "cycleway:right" = 'track' OR
"cycleway:oneside" =  'lane' OR "cycleway:otherside" = 'lane' OR
"path.bicycle" = 'designated'
```

2.2 Ordnance Survey data

The Ordnance Survey (OS) data was obtained from their central office following completion of a non-disclosure aggreement. Following substantial software challenges, requiring expensive propriatary software, the entire OS dataset was loaded for the UK and could be analysed.

The full attributes of the OS data will be described in the final paper, as will the methods, and results. Preliminary analysis suggests that where OS and OSM datasets do overlap (in a minority of the paths' lengths for both datasets), the correspondence between them is good, with the OS dataset having higher spatial resolution (Figure 2).

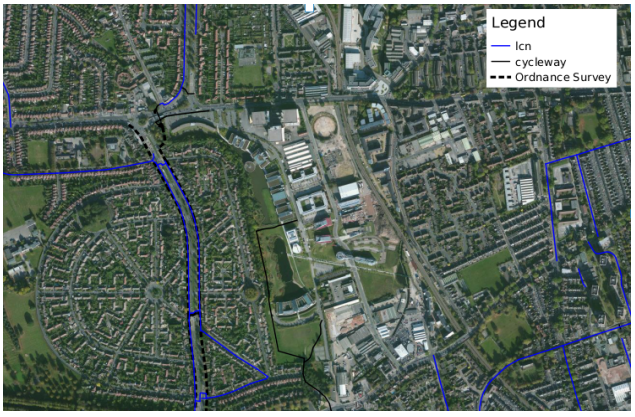

Figure 2: Example of partial correspondence between OSM and OS cycle path data. Note this is not at all representative of the UK.

3 Conclusion

This is the first study to our knowledge which decisively finds a free and crowd-sourced dataset to be more appropriate to a critical application than a professionally produced proprietary dataset from a national cartographic organisation. In this case, Open Street Map was found to provide more coverage, attributes, continuity and recency than the commercial dataset provided by the UK's national geospatial data provider Ordnance Survey.

Many future research directions are opened-up by this study, including more systematic and 'real time' tests of different geographical datasets to provide 'health checks' of suitability for different applications; other niche areas where crowd sourced geographical data may be more appropriate than centralised sources; and the potential for national mapping bodies such as Ordnance Survey to incorporate the richness of crowd-sourced offerings into their official products. All of these themes intersect with wider questions about the 'democratisation' of academia (Berry and Moss, 2006) and open up exiting possibilities for combining educational benefits with improved public data administration.

4 References

Arnold, Lisa L., and Paul A. Zandbergen. 2011. "Positional Accuracy of the Wide Area Augmentation System in Consumer-Grade GPS Units." *Computers & Geosciences* 37 (7) (July): 883–892. doi:10.1016/j.cageo.2010.12.011. `http://www.sciencedirect.com/science/article/pii/S0098300411001063http://www.sciencedirect.com/science/article/pii/S0098300411001063/pdfft?md5=a6e3fa02e5aa8a8208c04a9533764729/&pid=1-s2.0-S0098300411001063-main.pdf`.

Berry, D. M., & Moss, G. (2006). Free and open-source software: Opening and democratising e-government's black box. *Information Polity*, 11(1), 21–34.

Goodchild, Michael F. 2007. "Citizens as Sensors: the World of Volunteered Geography." *GeoJournal* 69 (4) (November): 211–221. doi:10.1007/s10708-007-9111-y. `http://link.springer.com/10.1007/s10708-007-9111-y`.

Goodchild, Michael F., and Sucharita Gopal. 1989. *The Accuracy Of Spatial Databases*. CRC Press. `http://books.google.co.uk/books?id=HL2O6J-XtLAC`.

Hunter, Gary J, and Michael F Goodchild. 1995. "Dealing with Error in a Spatial Database: A Simple Case Study." *Photogrammetric Engineering and Remote Sensing* 61 (5): 529–537.

Roller, D. 2010. *Eratosthenes' "Geography"*. Princeton University Press. `http://books.google.co.uk/books?id=8peKyWK/_SWsC`.

Thapa, Khagendra, and John Bossler. 1992. "Accuracy of Spatial Data Used in Geographic Information Systems." *Photogrammetric Engineering and Remote Sensing* 58 (6): 835–841.