

# When worlds collide: combining Ordnance Survey and Open Street Map data

**Suchith Anand\***, **Jeremy Morley\***, **Wenchao Jiang\***, **Heshan Du\***, **Glen Hart#**, **Mike Jackson\***

\*Centre for Geospatial Science, University of Nottingham

#Ordnance Survey, UK

---

## Introduction

The context of this paper is the progress of national and international spatial data infrastructures such as the UK Location Programme and INSPIRE, contrasted against crowd-sourced geospatial databases such as Open Street Map. While initiatives such as INSPIRE tend towards a top-down process of harmonised data models and services using ISO & OGC standards, the OSM approach is one of tagged data with attribute tags agreed through consensus, but a tag set that can change with time (with inherent related issues of data quality). There is a danger that should the more formal approaches simply ignore the crowd sourced initiatives then they will miss an opportunity to evolve to better meet growing demands for geographic information. In any case both formal and informal data will increasingly coexist begging the question of how an end user gains maximum benefit from both.

Ordnance Survey as the national mapping agency of Great Britain provides authoritative datasets with published data specifications driven by a combination of user need and the history of national mapping with a remit to ensure real-world feature changes are reflected in the OS large-scale data within 6 months. OSM in contrast relies on the availability of local mapping enthusiasts to capture changes but through its more informal structure can capture a broader range of features of interest to different sub-communities such as cyclists or horse riders.

This research has been carried out to understand the issues of data integration between crowd sourced information and authoritative data. The aim of the research was to look into the mid-term and long-term effects of crowd sourcing technologies for understanding their effects on the change intelligence operations of national mapping agencies (NMAs) in the future. Mobile phones, with more computing power than the desktop machine of 5 years ago and incorporating built-in GPS receivers and cameras have become widespread and give people a multi-sensor capability. This combined with CCTV, sensor webs, RFID etc. offers the potential to make data capture pervasive and ubiquitous. All key sectors of modern economies will be affected by the developments in crowd sourcing of information. The synergies created by new technologies will create the conditions for exciting new developments in geospatial data integration. This has an impact in the spatial data collection domain especially in collecting vernacular and crowd-sourced information. Individual users will be able to use these technologies to collect location data and make it available for multiple applications without needing prior geospatial skills.

The basic question behind our research is how do we combine data from authoritative OS data sets with feature-rich, informal OSM data, recognising the variable coverage of OSM while capturing the best of both worlds? There have been previous studies (Al-Bakri and Fairbairn, 2010) focussing on geometric accuracy assessment of crowd-sourced data(OSM) with OS data.

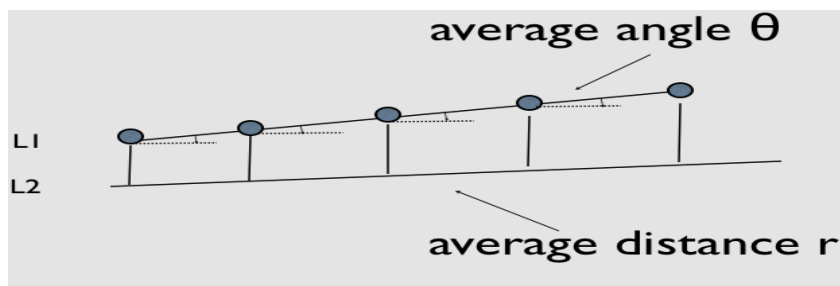
Another important context is the rapid developments in Open Source GIS. The availability of free and open source GIS has made possible for large number of government organizations and SMEs to make use of GIS tools in their work. The Open Source Geospatial Foundation (OSGeo) is an excellent example of community initiative to support and promote the collaborative development of open geospatial technologies. OSGeo's key mission is to promote the use of open source software in the geospatial industry and to encourage the implementation of open standards and standards based interoperability in its projects.

## Methodology

As part of this research, two techniques were developed for integrating authoritative data with crowd sourced data based on map matching (described in this paper) and ontology based matching. Automated map-matching is a fundamental research area in GIS. Map-matching algorithms integrate positioning data with spatial road network data (roadway centrelines) to identify the correct link on which a vehicle is travelling and to determine the location of a vehicle on a link (Quddus et al., 2007). Map matching also uses vector map information integrating various positioning sensor data to produce the best estimate of vehicle position (Xi L et al, 2007). Map matching techniques can be adapted to the process of data conflation. This paper introduces the concepts of this positional matching algorithm, and looks into how this map matching technique can be applied to combine authoritative data and crowd sourced data. Results of experimentations are presented and an evaluation is conducted to assess the effectiveness of the algorithm in this data conflation context.

For data conflation, the first challenge is how to find correspondence in two datasets, that is, recognizing features that represent the same object in the real world. Geometry information of two datasets is not perfectly aligned. The first step of the combination is to identify correspondent features in two datasets using the adapted automated map matching algorithms.

Although the situation is similar, position matching cannot be transposed to the process of data conflation directly. In order to apply position matching to data conflation, the nodes are extracted from the line features in one dataset at given distances along the features. Then, for every node, values for average distance  $r$  and average angle  $\theta$  are calculated to each candidate reference line feature.  $\lambda$  for every candidate is then calculated as per the following equation;  $\lambda = \omega_1 \times \text{average}(r) + \omega_2 \times \text{average}(\theta)$ .



**Fig 1- Example of applied map matching approach**

There are two criteria for the matching process: (i) firstly  $\lambda$  should be less than the threshold value; (ii) then the feature candidate with the lowest  $\lambda$  is matched. The pseudo code below describes the process for the Position Matching algorithm implemented.

```

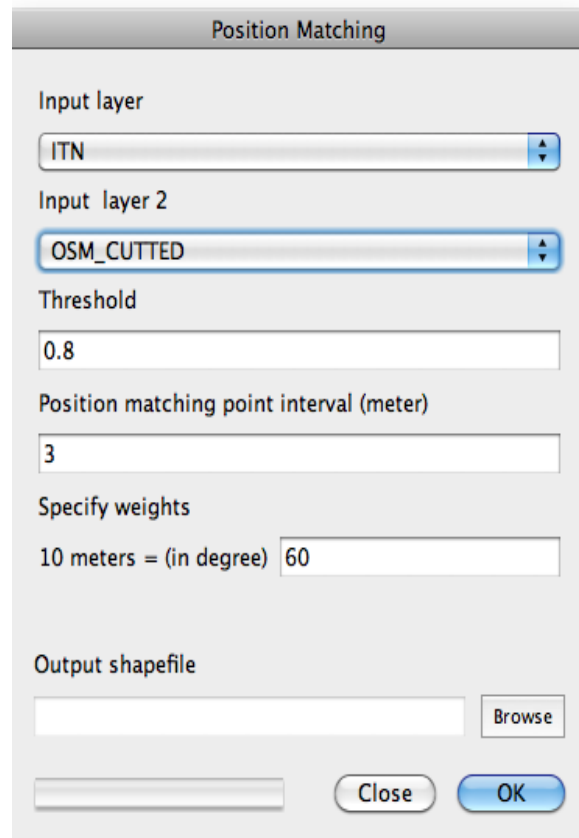
Function PositionMatch(inputFeature,candidateSet,threshold );
input:
  inputFeature: a feature in one dataset
  candidateSet: candidates in the other dataset
  threshold
output: matched feature chosen from candidates

for each inputFeature in dataset1
  nodeset←extractNode(feature)
  for each candidate in candidateSet
    for each node in nodeset
      r←getR(candidate,node)
      θ←getθ(candidate,node)
      list.add(r,θ)
    λ ← calculate_λ(list)
    if λ < threshold
      bestCandidate←UpdateCandidateWithLowest_λ(candidate, λ)
return bestCandidate

```

## Implementation

Ordnance Survey Integrated Transport Network (ITN) data and OpenStreetMap (OSM) road data for Portsmouth, UK were used as a case study to explore methodologies to integrating two heterogeneous data sources. **Geometry information is not completely aligned for the datasets. For the case study area there are 565 features in ITN and 479 features in OSM dataset.**



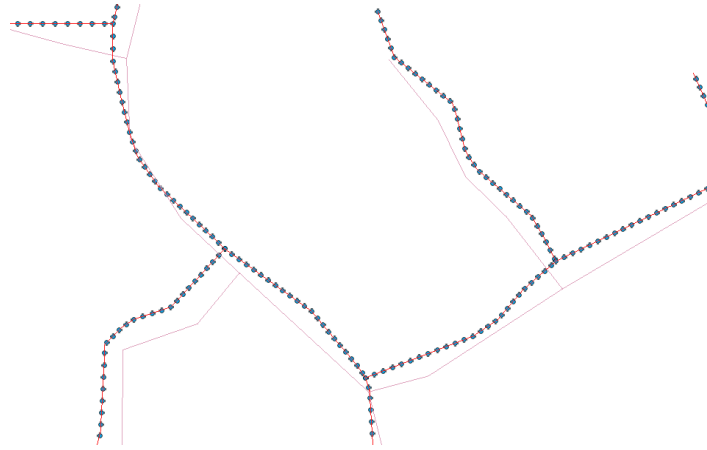
The screenshot shows a dialog box titled "Position Matching". It contains the following elements:

- Input layer:** A dropdown menu with "ITN" selected.
- Input layer 2:** A dropdown menu with "OSM\_CUTTED" selected.
- Threshold:** A text input field containing "0.8".
- Position matching point interval (meter):** A text input field containing "3".
- Specify weights:** A section with a label "10 meters = (in degree)" and a text input field containing "60".
- Output shapefile:** A text input field with a "Browse" button to its right.
- Buttons:** "Close" and "OK" buttons at the bottom right.

**Fig 2. User interface of map matching tool**

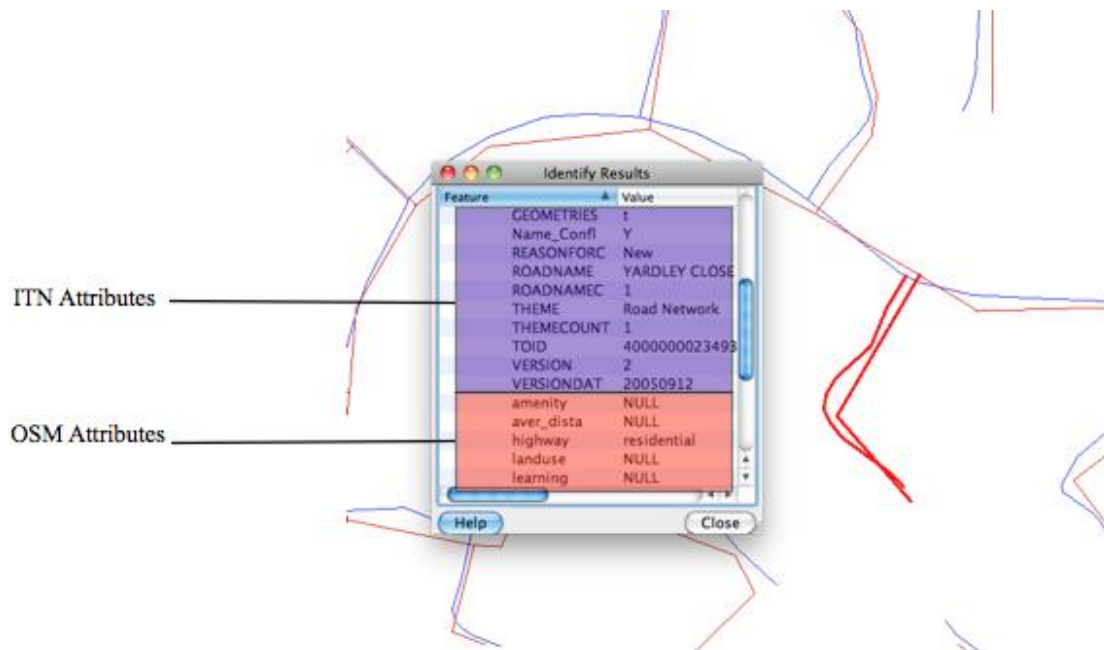
The algorithm has been implemented as a Python plugin in the open source Quantum GIS software. In the user interface (Fig. 2) of the Python plugin, the threshold for  $\lambda$ , the position matching interval and the relative weights (effectively  $\omega_1$  &  $\omega_1$ ) can be specified.

ITN data is taken as the reference system and features in the ITN are expanded into nodes every 3 meters along line features in order to calculate average  $\theta$  and average  $r$  to their candidates.



**Fig 3 . Nodes extracted from ITN features**

The output layer of the plugin is shown in Fig 4. Matching pairs of features are output to one layer, with their attribute information combined. It should be noted that the output layer is not simply an overlay of two maps. Correspondent features in two datasets are recognised and their attributes are merged (fig. 5).



**Fig 4 . Combined Layer (ITN in blue; OSM in red)**

### ITN Attributes

### OSM Attributes

Attribute table - 8\_60

	DESCRIPTO	DESCRIPT1	DESCRIPT2	ROADNAME	ROADNAMEC	GEOMETRIES	BOUNDEDBY	timestamp	user	tags	name	place	h
0	1	NULL	0	MARAZAN ...	1	t	466507.0,1...	2010-06-0...	stuphi	"highway"=...	Marazan Road	NULL	uncla
1	1	NULL	0	MARAZAN ...	1	t	466507.0,1...	2010-06-0...	stuphi	"highway"=...	Marazan Road	NULL	uncla
2	1	NULL	0	MARAZAN ...	1	t	466507.0,1...	2010-06-0...	stuphi	"highway"=...	Marazan Road	NULL	uncla
3	1	NULL	0	MAYFIELD R...	1	t	465100.0,1...	2009-03-2...	nickw	"created_by...	Mayfield Road	NULL	residi
4	1	NULL	0	MAYFIELD R...	1	t	465100.0,1...	2009-03-2...	nickw	"created_by...	Mayfield Road	NULL	residi
5	1	NULL	0	EGAN CLOSE	1	t	465753.0,1...	2007-11-2...	paulo	"created_by...	Egan Close	NULL	residi
6	1	NULL	0	WILLIAMS R...	1	t	466490.69...	2010-06-0...	stuphi	"highway"=...	Williams Road	NULL	uncla
7	1	NULL	0	WILLIAMS R...	1	t	466490.69...	2010-05-3...	Wynndale	"highway"=...	NULL	NULL	uncla
8	1	NULL	0	WILLIAMS R...	1	t	466490.69...	2010-05-3...	Wynndale	"highway"=...	NULL	NULL	uncla
9	1	NULL	0	WEMBLEY G...	1	t	466463.0,1...	2009-03-1...	TimSC	"created_by...	Wembley Gr...	NULL	residi
10	1	NULL	0	WILLIAMS R...	1	t	466490.69...	2010-05-3...	Wynndale	"highway"=...	NULL	NULL	uncla
11	1	NULL	0	WEMBLEY G...	1	t	466463.0,1...	2009-03-1...	TimSC	"created_by...	Wembley Gr...	NULL	residi
12	1	NULL	0	WILLIAMS R...	1	t	466490.69...	2010-06-0...	stuphi	"highway"=...	Williams Road	NULL	uncla
13	1	NULL	0	WEMBLEY G...	1	t	466463.0,1...	2009-03-1...	TimSC	"created_by...	Wembley Gr...	NULL	residi
14	1	NULL	0	WILLIAMS R...	1	t	466490.69...	2010-05-3...	Wynndale	"highway"=...	NULL	NULL	uncla
15	1	NULL	0	WEMBLEY G...	1	t	466463.0,1...	2009-03-1...	TimSC	"created_by...	Wembley Gr...	NULL	residi
16	1	NULL	0	WILLIAMS R...	1	t	466490.69...	2010-06-0...	stuphi	"highway"=...	Williams Road	NULL	uncla
17	1	NULL	0	LARKHILL R...	1	t	465851.0,1...	2007-11-2...	paulo	"created_by...	Larkhill Road	NULL	residi
18	1	NULL	0	MERLIN DRIVE	1	t	466012.0,1...	2009-08-0...	Milliams	"highway"=...	Merlin Drive	NULL	residi
19	1	NULL	0	MERLIN DRIVE	1	t	466012.0,1...	2009-08-0...	Milliams	"highway"=...	Merlin Drive	NULL	residi
20	1	NULL	0	MERLIN DRIVE	1	t	466012.0,1...	2009-08-0...	Milliams	"highway"=...	Merlin Drive	NULL	residi
21	1	NULL	0	HOBBY CLOSE	1	t	466027.0,1...	2009-03-1...	TimSC	"created_by...	Hobby Close	NULL	residi
22	1	NULL	0	BENHAM D...	1	t	465871.0,1...	2009-03-1...	TimSC	"created_by...	Benham Drive	NULL	residi
23	1	NULL	0	BREECH CL...	1	t	465865.0,1...	2009-03-1...	TimSC	"created_by...	Breech Close	NULL	residi
24	1	NULL	0	BREECH CL...	1	t	465865.0,1...	2009-03-1...	TimSC	"created_by...	Breech Close	NULL	residi
25	1	NULL	0	BREECH CL...	1	t	465865.0,1...	2009-03-1...	TimSC	"created_by...	Breech Close	NULL	residi
26	1	NULL	0	HONEYWO...	1	t	465905.11...	2009-03-1...	TimSC	"created_by...	Honeywood...	NULL	residi
27	1	NULL	0	HONEYWO...	1	t	465905.11...	2009-03-1...	TimSC	"created_by...	Honeywood...	NULL	residi
28	1	NULL	0	HONEYWO...	1	t	465905.11...	2009-03-1...	TimSC	"created_by...	Honeywood...	NULL	residi

Fig 5. Combined Attribute table (ITN and OSM)

Threshold	matching_features	percentage	distribution
0.1	21	4%	21
0.2	95	17%	74
0.3	170	31%	75
0.4	245	45%	75
0.5	331	60%	86
0.6	378	69%	47
0.7	407	74%	29
0.8	429	78%	22
0.9	445	81%	16
1.0	455	83%	10

Table 1: Percentage matching features (ITN) for different threshold values

Total sample size	TRUE		FALSE	
	Positive	Negative	Positive	Negative
56	47	1	0	8

**Table 2: True matching features (ITN) for a small sample size**

Analysis was carried out to understand the percentage of matching features in ITN for the different threshold values. The results are presented in Table 1. It can be seen that higher threshold value leads to a larger number of matching features. For example, when Threshold is set to 0.9, 445 features in the ITN find a correspondence in the OSM data set, which is 83% of the total features in sample ITN used for this study. A subset of the full dataset was used to check the true positive matching of the matched features in the two datasets. It was found (Table 2) that for this sample there was an 86% accuracy level with the current implementation, i.e. 48/56 features were correctly identified with their matches or not. Eight matches were not detected.

## Conclusions and Future Work

This paper looks into developing techniques for geospatial data integration using map matching techniques. There are very few examples of actual implementations and this study has been successful in developing tools to do this in one area. A prototype map matching technique has been developed to derive geometry based map matching between ITN<sup>®</sup> and OSM road datasets. Initial analysis of the sample data used shows good accuracy levels with the current implementation but more work needs to be done in refining the process. Future work will concentrate on refining the technique through the use of additional constraints to enhance visualization and usability, and to assess the quality and benefits of the merged dataset. The next stage in development will be to select the preferred geometry (from ITN or OSM, or a merger) for each matched feature pair. In addition, some criteria will be needed to assess unmatched features to decide whether the unmatched features are, for example, missing or deleted from the reference database.

NMAs can benefit immensely from such developments but research is needed to understand how to tap into this huge potential opportunity and to obtain a consistent, quality and verifiable product from the data so acquired within the terms of use of the crowd-sourced data. This can then be used to develop different models for example for change intelligence operations. Also there will be scope for deriving products based on the volunteered and vernacular geographic data collected from the crowd sourced communities. The software developed does show promising results when applied to the case study datasets. The software developed will help in further refining the integration of crowd sourced data with authoritative data. This when combined with ontology based attribute matching techniques offer promising results in the holistic integration of these disparate datasets.

## Acknowledgement

The authors express thanks for the Ordnance Survey, UK and OSM for the data used in this work. All figures in this text using OS data are ©Crown Copyright/database right 2010, an Ordnance Survey/EDINA supplied service. In addition we thank the Ordnance Survey for funding for Dr. Anand's work through the Future Data project.

## References

- Al-Bakri M, Fairbairn D., Assessing the accuracy of crowdsourced data and its integration with official spatial datasets, Proceedings of Accuracy 2010 Symposium, UK, 2010
- Quddus, M.A., Ochieng W.Y. Noland R.B., Current map-matching algorithms for transport applications: State-of-the-art and future re-search directions. Elsevier Transportation Research Part C, 15, pp.312–328, 2007.

Xi L, Liu Q, Li M, Liu Z., Map Matching Algorithm and Its Application, Advances in Intelligent Systems Research, ISKE-2007 Proceedings, 2007