Building Orientations Using Angle Clustering

Sheng Zhou[*] and Jonathan Simmons[†]

Data Office, Ordnance Survey, Great Britain

March 12, 2018

Summary

This study proposes an approach to determine the orientations of building polygons by clustering polygon edge vectors using weighted directional statistics. It can preserve orientation information of multi-orientated buildings. A perturbation technique is also introduced to handle boundary effect of modulo operation.

KEYWORDS: Building Orientation; Clustering; Machine Learning; Map Generalisation; Spatial Data Analysis

1. Introduction

Buildings is a prominent area feature category of many large-scale maps. Building orientation is an important descriptor of building characteristics and has numerous applications in cartography and other fields (e.g. generalisation of buildings (Lokhat and Touya 2016) and other map features, and feature group pattern recognitions (Yan, Weibel and Yang 2008; Zhang et al. 2013)).

Most buildings have straight walls aligned to only a few directions. Therefore, orientation descriptors for generic polygons (for example direction of the longest diagonal) may not be appropriate and more specific descriptors are required. Duchêne et al. (2003) evaluated five building orientation descriptors and proposed a new measure of "wall statistical weighting orientation" as the follows:

- For all edge vector e_i (p_{is}, p_{ie}) of length L_i = Distance(p_{is}, p_{ie}) of the polygon depicting the boundary wall of a building, convert the orientation α_i = Angle(p_{is}, p_{ie}) into α'_i $[0, \pi/2)$ via a $\pi/2$ modulo operation;
- For candidate orientation $\alpha \in [0, \pi/2)$ and angle tolerance δ, compute the contribution of edge e_i with converted orientation $\alpha'_i \in [0, \pi/2]$ as:
  - $C_i = 0$ (if Abs $(\alpha - \alpha'_i) > \delta$), or otherwise
  - $C_i = L_i * (1 - \text{Abs}(\alpha - \alpha'_i) / \delta)$
- Confidence indicator for orientation α: $CF_\alpha = \Sigma(C_i) / \Sigma(L_i)$
- Different candidate orientations in $[0, \pi/2)$ at a predefined step *s* are examined and the candidate with largest confidence factor is returned as the orientation of the building

This approach is very effective in finding the primary orientation of buildings and has since been adapted in many applications (e.g. Zhang et al. 2013). Nevertheless, there are a few issues to be addressed. The $\pi/2$ modulo is not always appropriate if the two main orientations are at right or obtuse angle. The modulo operation itself could be problematic in certain cases. For example, for two angles $a_1 = \pi/2 - \varepsilon$ and $a_2 = \pi/2 + \varepsilon$ (where $\varepsilon < \delta$) which are close to each other and near the modulo boundary, the $\pi/2$ modulo will leave a_1 at $\pi/2 - \varepsilon$ and convert a_2 to ε, which are almost $\pi/2$ apart.

[*] Sheng.Zhou@os.uk

[†] Jonathan.Simmons@os.uk

In this study we take a different view and treat building orientation as a problem of clustering edge vectors. We also propose a perturbation-based method to address the modulo boundary issue.

2. Clustering of Angle Values with Weights

2.1. Clustering building edge segments and directional statistics

Clustering is the task of grouping similar objects together and separating dis-similar objects into different clusters. It is commonly used in many fields. In the context of machine learning, clustering is normally regarded as an un-supervised learning process.

There are many generic clustering algorithms applicable to a wide range of problems. However, edge vector clustering will require some special treatments to handle orientation values using concepts in circular data analysis and directional statistics (Maridia and Jupp 2000).

If we treat an edge as a vector with its origin at (0,0), it may be referenced by a 2D point (α_i, L_i) in polar coordinate system where $\alpha_i \in (-\pi, \pi]$ and $L_i \in (0, \infty)$. Obviously, it makes no sense to cluster these points directly using conventional cluster algorithms. Instead, we view the points as on the unit circle with the length as the weight w_i. Consequently, the weighted mean θ_r for a collection of n angle references A = {$a_i(\alpha_i, \mid w_i = L_i) \mid i = 1, n$} may be defined as:

$$X = \frac{\sum_1^n w_i * \cos\alpha_i}{\sum_1^n w_i}$$
$$Y = \frac{\sum_1^n w_i * \sin\alpha_i}{\sum_1^n w_i}$$
$$\bar{R} = \sqrt{X^2 + Y^2}$$
$$\theta_r = atan2(Y, X)$$

$\boldsymbol{V = 1 - \bar{R}}$ in range [0, 1] is called **circular variance** which is a measure of dispersion of the angle collection.

2.2. Agglomerative hierarchical clustering for weighted angle values

Weighted angle values representing building wall edge orientations may be clustered in a bottom-up hierarchical manner:

- If required, apply modulo operation (π or π /2) to convert angle values in angle references
- Merge angle refences with the same angle value into a single reference and the sum of weights as new weight
- Generate a cluster for each angle reference
- While the termination criteria **T** are not met
    - Select the pair of closest clusters by distance metric **D** and merge them into one cluster
    - repeat

There are many options for the metric **D** (e.g. minimum circular difference between two clusters). Currently we use the resulting circular variance if two clusters were merged as the metric. In each iteration we merge the pair with the smallest intra-cluster variance increase.

It is always a challenge to find the proper termination criteria for any clustering process. In view that initially all variance are inter-cluster and that intra-cluster variance will increase as clusters are merged, we used the maximum ratio of intra-cluster variance over initial total variance as the termination criterion for most experimental results presented here. Alternatively, the maximum angular width of clusters may be used. It should be noted these by no means are the best options.

2.3. A perturbation approach for better modulo operation

In section 1 we mentioned the boundary effect in modulo operation. This issue may easily be addressed by performing a second round of clustering, in which a perturbance (rotating by a given angle) is applied to the angle values before modulo operation. A reverse rotation will be applied to the result afterwards.

Of the two results, the better one (fewer clusters, or less disperse, or by some other measures) will be selected as the output. For more irregular polygons, further rounds of perturbation may be required.

This technique may also be used in the original wall statistical weighting orientation algorithm to improve its sensitivity.

3. Implementation and Experiments

The above algorithm has implemented in Java. Source code are available at GitHub (https://github.com/szhou68/AngleClustering).

Initially the implementation was tested on a manually generated data set (**Figure 1**). Polygons in the columns C and D contain vertical and horizontal edges. Columns A/B and E/F are the results of a small left and right tilt applied to C/D respectively.

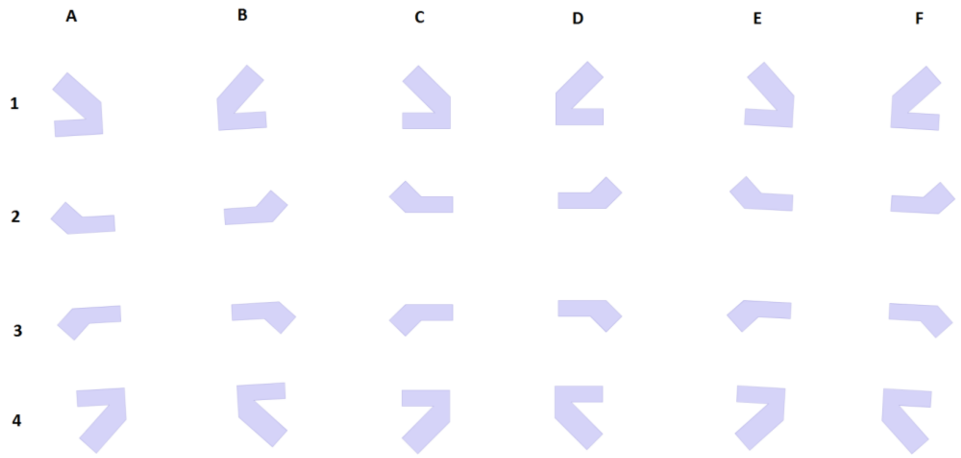

Figure 1 Manually generated test dataset

Figure 2 shows the clusters generated without modulo conversion at intra-cluster variance ratio threshold of 0.01. Clusters are represented by lines showing the sum of cluster member weights projected on the mean orientation of the cluster.

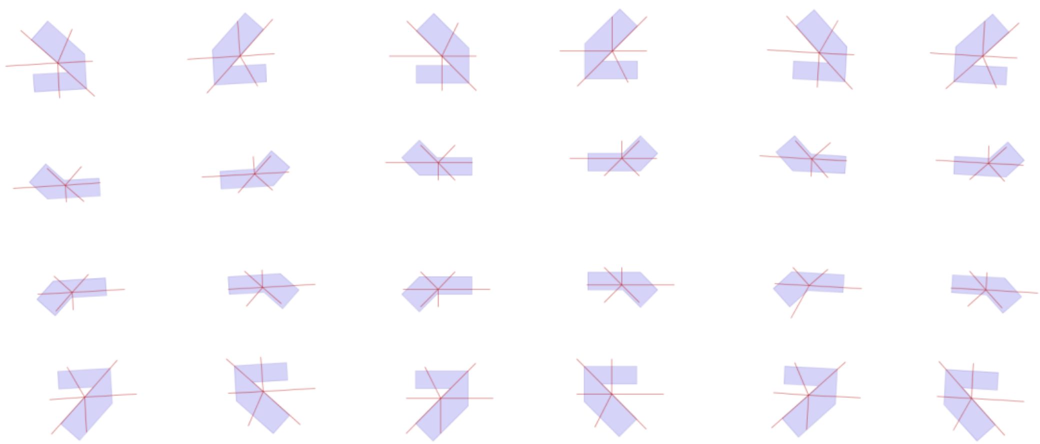

Figure 2 Clusters without modulo operation

Figure 3 are results with π-modulo. Note in columns C and D the results without perturbation contain 5 clusters instead of 4. With perturbation, two collinear clusters are merged.

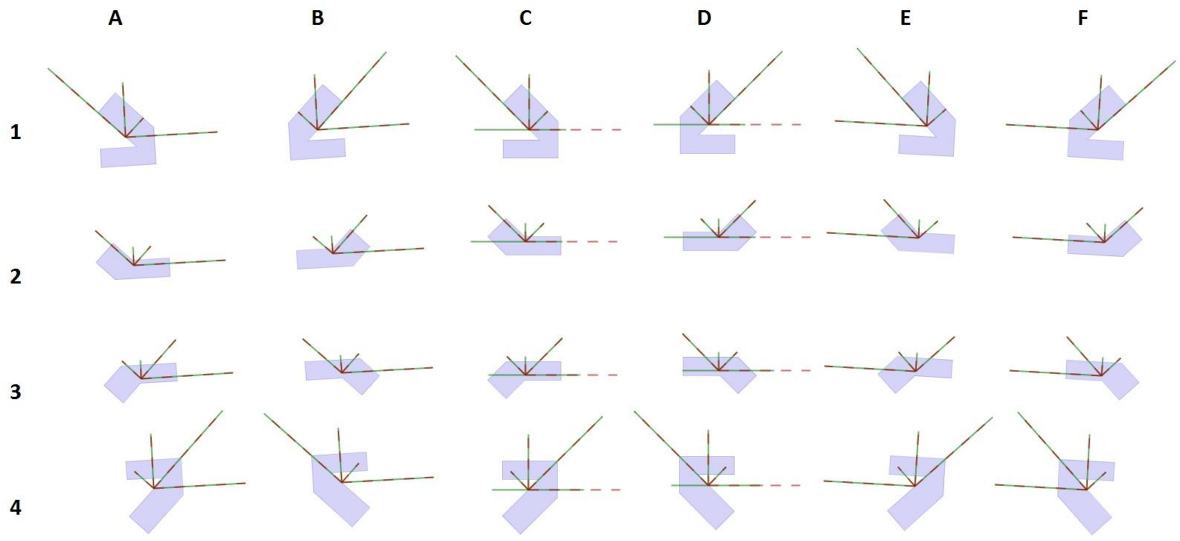

Figure 3 π-modulo clusters with (dash) and without (solid) π/12 perturbance

Figure 4 are results with π/2-modulo. Here Perturbation also helps to merge two perpendicular clusters (Columns C and D).

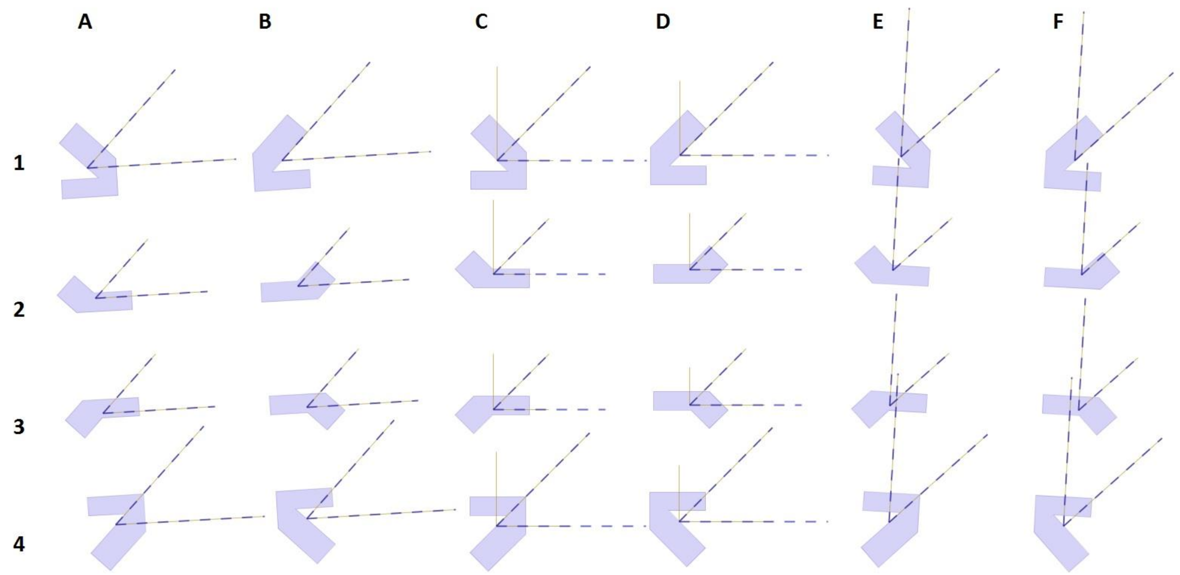

Figure 4 π/2-modulo clusters with (dash) and without (solid) π/12 perturbance

Figure 5 compares results of π-modulo to π/2-modulo. |It appears that π-modulo preserves more orientation information of buildings. The π/2-modulo converses orientations into [0, π/2) quadrant. Therefore, in case a building (e.g. A1, C1 and E1) has a major orientation along quadrants [π/2, π) and [-π/2, 0), a perpendicular orientation is returned instead. This is not an issue for π-modulo.

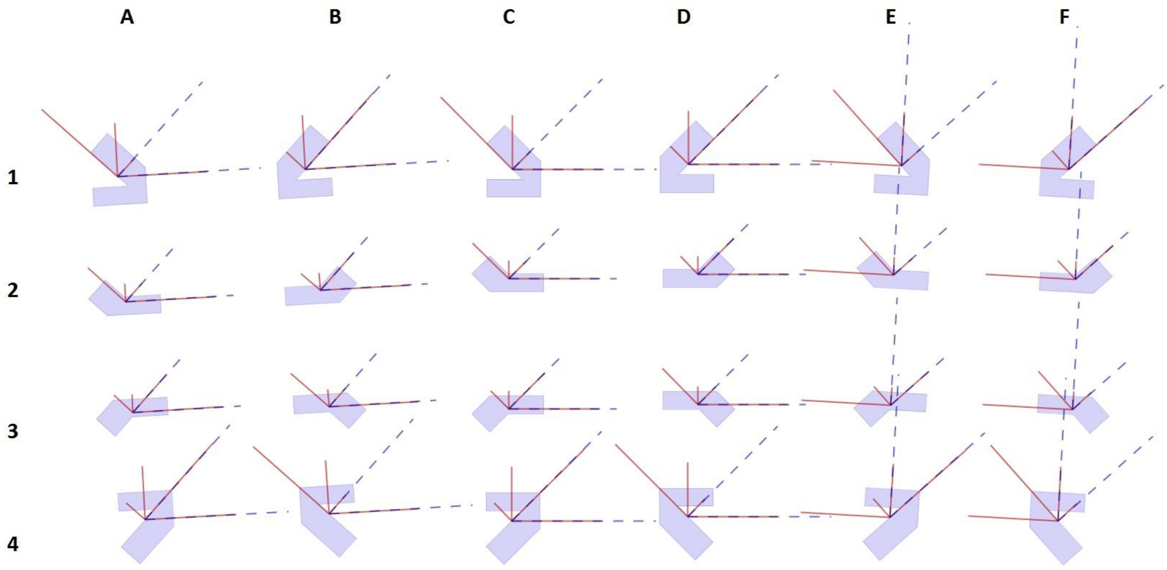

Figure 5 π-modulo (solid line) vs π/2-modulo (dash-line) with perturbance

Figure 6 shows some more complicated examples from OS MasterMap topographic layer.

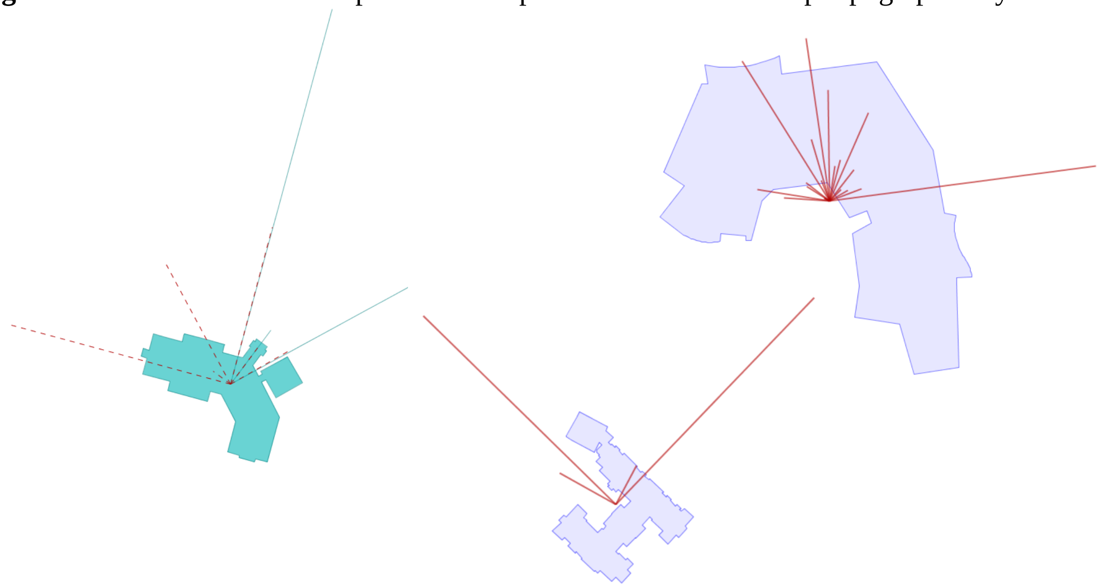

Figure 6 OS MasterMap Building Polygon examples (Ordnance Survey © **Crown copyright** 2018)

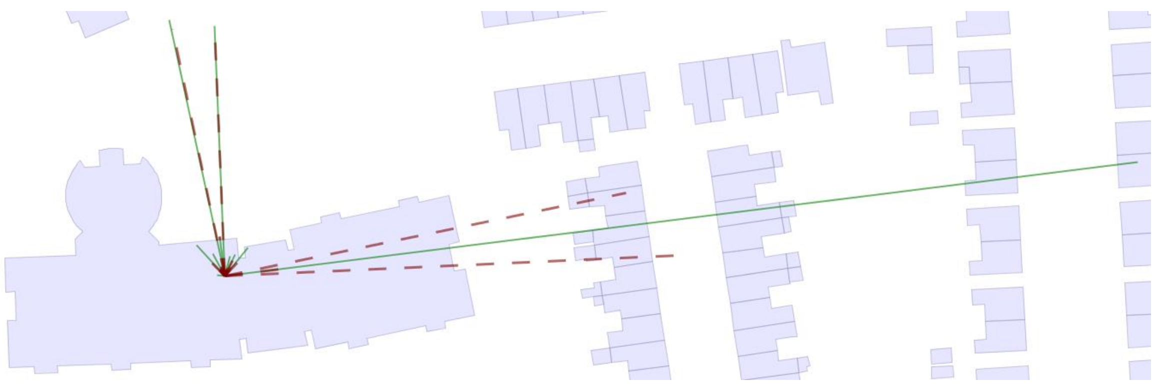

Figure 7 Variance ratio vs angular width (Ordnance Survey © **Crown copyright** 2018)

Figure 7 compares the results of the two termination criteria we tested. Results using maximum intra-cluster variance ratio are shown in dash line and the maximum angular width (15 degree) results are in solid line.

4. Discussion

Our initial experiment results indicate that the clustering approach has the potential to detect multiple orientations of buildings effectively. The perturbation method offers a solution to the boundary effect issue in modulo operations. π-modulo seems to be the option if applications require information on two or more major orientations of a building.

It remains an issue to determine the optimal number of clusters that should be generated from edge orientations. We briefly examined the conventional "elbow" and silhouette methods but it appears that while they seem to work well on large buildings with more edge segments, they don't warranty consistent results, probably due to the small sample size and skewed value distribution of building edge orientations.

The variance ratio we used in our initial experiment is sensitive to data and even sensitive to the modulo operation (which hugely reduces variance). At present we believe the best option might be the maximum angular width of cluster (and probably in combination with minimum angular distance between clusters), which generates clearly more interpretable results. We aim at exploring these four methods in depth in future studies.

5. Biography

Sheng Zhou is a Senior Data Scientist at Ordnance Survey. His research interests include: application of machine learning techniques on spatial data; computational geometry; spatial databases; spatial analysis; map generalisation.

Jonathan Simmons is a Principal Consultant at Ordnance Survey and the lead of Data Science and Analytics team at the Data Office of Ordnance Survey.

References

Duchêne C, Bard S, Barillot X, Ruas A, Trévisan J and Holzapfel F (2003). Quantitative and qualitative description of building orientation. *ICA Workshop on Progress in Automated Map Generalization*, 28-30 April 2003, Paris (France).

Lokhat I and Touya G (2016). Enhancing building footprints with squaring operations. *Journal of Spatial Information Science*, 13, 33-60

Maridia KV and Jupp PE (2000). *Directional Statistics*. John Wiley & Sons Ltd.

Touya G, Girres J (2014). Generalising Unusual Map Themes from OpenStreetMap. *17th ICA Workshop on Generalisation and Multiple Representation*, 17 September 2014, Vienna (Austria).

Yan H, Weibel R and Yang B (2008). A Multi-parameter Approach to Automated Building Grouping and Generalization. *Geoinformatica*, 12, 73–89

Zhang X, Stoter J, Ai T, Kraak M, and Molenaar M (2013). Automated evaluation of building alignments in generalized maps. *International Journal of Geographical Information Science*, 27(8), 1550-1571.