Published in: Kyriakidis, P., Hadjimitsis, D., Skarlatos, D. & Mansourian, A. (eds.), 2019. Accepted Short Papers and Posters from the 22nd AGILE Conference on Geo-information Science. Cyprus University of Technology 17-20 June 2019, Limassol, Cyprus. ISBN 978-90-816960-9-8

Determining Point Locations of Populated Places by Using Area Datasets

Thomas Brinkhoff
Jade University Wilhelmshaven/Oldenburg/Elsfleth
Institute for Applied Photogrammetry and Geoinformatics (IAPG)
Ofener Str. 16/19, D-26121 Oldenburg, Germany
thomas.brinkhoff@jade-hs.de

Abstract

For mapping applications, we often illustrate the position of populated places like towns, villages and hamlets by point symbols. However, suitable datasets of such point locations are often missing or have bad or unpredictable quality. This leads to unclear or confusing maps. In contrast, statistical institutions often provide area datasets for such places. Furthermore, datasets of built-up areas are available or can be computed by using open data. In this paper, the question is investigated how such datasets can be used for deriving suitable point locations for populated places. An algorithm is presented and the impact of preprocessing steps on the quality of the point coordinates is investigated. First applications of the method show promising results.

1 Introduction

An essential task for designing meaningful and appealing maps is an adequate symbolizing of relevant map features. Almost all web applications indicate populated places like towns, villages and hamlets by point symbols. Figure 1 shows Google Earth as an example. The map visualizes a part of India. The place symbols refer to villages. However, their placement does not coincide with the depicted settlement areas. Thus, these symbols are unclear and confuse any user. It is impossible to decide to which settlement a symbol belongs.

Figure 1. Place symbols for villages in India by Google Earth (© 2016 CNES/Airbus, Google).

The depicted map is quite typical because high-quality point datasets of populated places are missing for many countries. There are many global open datasets with place coordinates. As we discuss later in detail, their quality is often bad or unpredictable. Metadata indicating the original source and accuracy of the coordinates are typically missing.

In contrast, for many countries area data for populated places are available. Statistical organizations often provide such *place areas*. They need them for organizing and evaluating surveys like population censuses. A typical example are the well-known TIGER files of the U.S. Census Bureau. The quality of such datasets is assured by official authorities and is sufficient for most applications. A straightforward solution would be to compute the centroids of those areas and to use them as place location. However, such an approach leads to a bad symbol placement when the areas include also unsettled areas.

In the last few years, many datasets of *built-up areas* became available. A user can use such a dataset directly or can use it for deriving built-up areas. The basic idea is to intersect such land coverages with place areas and then to compute suitable point locations. Such an approach will at least guarantee that the point symbol is located over a built-up area. The investigation in the following of this paper will however show that further processing steps are suitable in order to receive better results.

The rest of the paper is organized as follows: In the next section, we discuss the availability and shortcomings of global point datasets for populated places. Datasets suitable for extracting built-up areas are also topic of that section. Section 3 presents the algorithmic steps for computing point locations of populated places. A first evaluation of the results follows in the fourth section. The paper concludes with a short summary and an outlook to future work.

2 Global Datasets

The following discussion will cover only global datasets that allow a general use for (web) map applications. For some countries or regions, more suitable datasets or datasets with higher quality may exist.

2.1 Point Datasets for Populated Places

Three global sources for place locations are often used or citied: GeoNames, OpenStreetMap and Wikipedia.

GeoNames (http://www.geonames.org/) provides (as of February 2019) WGS 84 point coordinates for about 4.8 million populated places. Deficits include missing national identifiers, incomplete administrative information and missing quality indicators. According to our experience, the accuracy of coordinates is often (unpredictable) poor.

OpenStreetMap (OSM; http://www.openstreetmap.org) contains – among many other information – also place locations. They are represented by node elements that are marked with the key "place" (4.4 million, as of February 2019). In addition to the WGS84 coordinates, such elements may provide information about the administrative level and the official place identifier. Quantitative quality indicators for location information are missing. Furthermore, the actual allocation of these attributes and the accuracy of place coordinates vary.

Wikipedia provides for many articles a point coordinate that a user can extract by using different tools. Most practical is often the use of WikiData (https://www.wikidata.org/). If the corresponding information is stored in WikiData, the assignment of national identifiers to WikiData keys can be determined with the help of a SPARQL query. These keys allow retrieving coordinates. Information about origin and accuracy of these coordinates, however, is mostly missing.

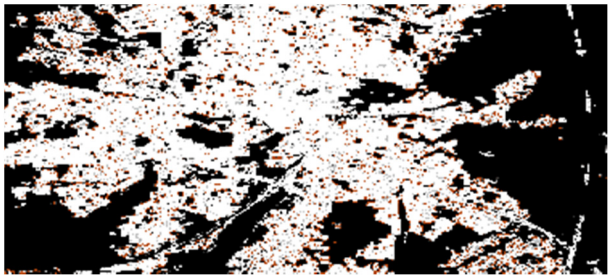
Overall, none of the mentioned datasets is suitable for a quality-assured symbolizing of populated places.

2.2 Datasets for Built-up Areas

If required on global scale, information about built-up areas will be typically derived from remote sensing data. Some few examples are the usage of Landsat 8 Operational Land Imager (OLI) data (Bhatti & Tripathi, 2014), the extraction from Advanced Spaceborne Thermal Emission and Reflection (ASTER) radiometer data (Miyazaki et al., 2014), and the use of Landsat TM/ETM+ images (Zhang et al., 2014).

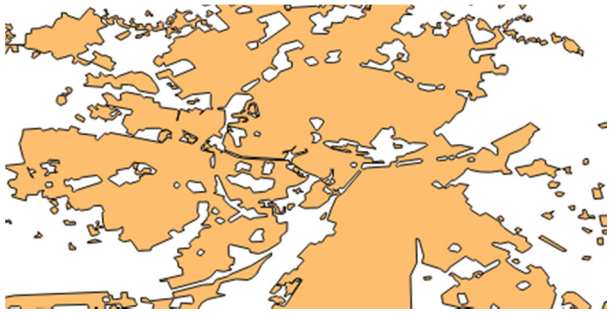
The currently most relevant data set in this context is the "Global Urban Footprint" (GUF), which was derived from radar data (Esch et al. 2013). About 180,000 individual images from the years 2010 to 2013 by the two radar satellites TerraSAR-X and TanDEM-X were processed and analyzed. As results, the earth's surface is subdivided into populated and unpopulated areas. At the beginning of 2017, DLR provided the resulting datasets in two resolutions: 12m for scientific use (Figure 2 shows an example) and 84m non-commercial use free of charge (http://www.dlr.de/guf/).

Figure 2. Build-up areas (white) of the City of Oldenburg (Germany) from the GUF dataset (12m).



Some studies have investigated deriving land use from OpenStreetMap data. This has been done mostly in local scale, typically for one or several cities (Vaz & Jokar Arsanjani, 2015) (Jokar Arsanjani et al. 2015), for smaller regions (Dorn et al., 2015) or single countries (Estima & Painho, 2013). We presented an approach for using the OSM dataset on global scale for deriving built-up and urban areas (Brinkhoff 2016). Figure 3 shows an example of built-up areas derived from OSM.

Figure 3. Built-up areas of the City of Oldenburg (Germany) derived from OSM.



3 Computing Point Locations of Populated Places

As motivated in the introduction, starting point are datasets with built-up areas and place areas. The place areas do not overlap with each other. A corresponding algorithm can easily be sketched:

(1) The place areas are intersected with the built-up areas. In respect to one place area, three cases can occur: The resulting geometry (a) is empty, (b) is a polygon or (c) is a multi-polygon. In case (a), the algorithm fails for this place.

(2) The place coordinate is located appropriately in the (multi-)polygon computed by step 1.

Algorithmic solutions for step 2 are presented in the following subsection. In this simple form, the algorithm often computes inappropriate results for real data. Section 3.2 discusses possible improvements in more detail.

3.1 Computing the Visual Center

The computation of a polygon's centroid or the center of its minimum bounding box can be performed with linear effort but leads obviously to bad results. There are different approaches in the literature that determine exactly or approximatively the

"geodesic center" or "visual center" of a polygon. Such center can be characterized as the point in a polygon that minimizes the maximum internal distance to any point in the polygon. Examples are methods that calculate or approximate the center using Voronoi diagrams (Asano & Toussaint 1986), polygon skeletons (Pollack et al. 1989) or quadtrees (Agafonkin 2016).

Garcia-Castellanos & Lombardo (2007) proposed an approximate method, which calculates iteratively smaller areas. This approach can be easily realized by using an inverse buffer calculation (i.e. the calculation of a buffer zone with negative distance). If no area remains after an iteration step, the absolute distance must be reduced and the step is executed again. If the area of the resulting (multi-)polygon is smaller than a given threshold (e.g., 1% of the area of the initial (multi-)polygon), the processing ends and returns the centroid of the resulting polygon or of the largest component of the resulting multi-polygon. This approach is simple to implement, shows sufficient performance and its approximate nature fits to the intended use case. Figure 4 shows an example calculation in three iteration steps.

Figure 4. Computation of the visual center by iterative inverse buffering.

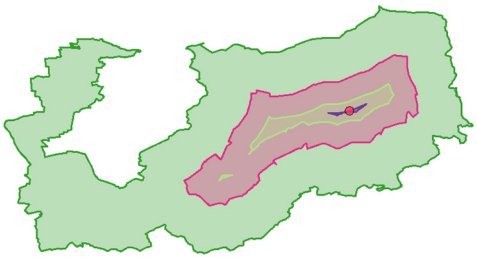
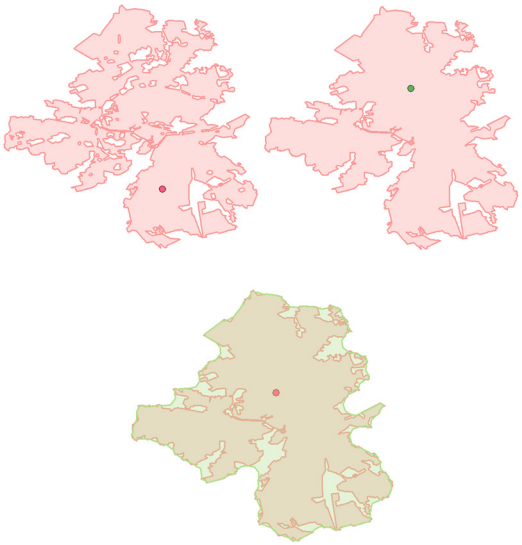


Figure 5. The calculation of the visual center on OSM data (top left), after removing small holes (top right) and after removing small holes and narrow indentations (bottom).

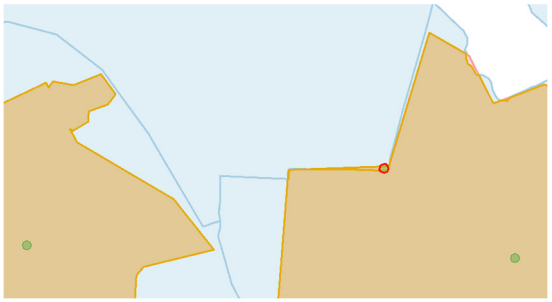


3.2 Data Preparation

Figures 2 and 3 show that potential built-up area datasets have a large number of holes and indentations. They outlast the intersection step, so that they are – without processing – included in the polygon for determining the place location. As shown in Figure 5 (top left), they impair the calculation. Therefore, small holes (Fig. 5, top right) and narrow indentations (Fig. 5, bottom) should be removed beforehand.

A further problem results from the fact that built-up areas as well as place boundaries are of limited accuracy. This property can falsify the result, especially in the case of locations whose built-up areas are missing in the dataset. In such cases, the built-up area of a neighbored place may overlap slightly the place polygon. Figure 6 shows an example. Therefore, the place areas should be reduced in size beforehand. This can be achieved by an inverse buffering with a distance that considers the inaccuracies of the datasets involved.

Figure 6. Without suitable preprocessing, the red dot would be incorrectly determined as suitable point location for the upper light blue place area.



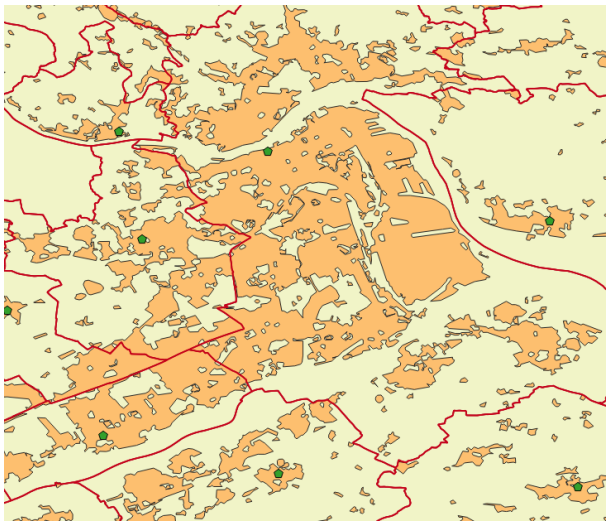
4 Evaluation

The evaluation of a suitable placement of point symbols for populated places is a difficult task. In the following, we will compare the computed positions with coordinates that are defined by an official institution. The results will obviously differ. Nevertheless, such comparison allows a first assessment of our approach as well as the investigation of impacts of varying parameters and algorithmic steps.

For this first evaluation, the following datasets are used (see also Figure 7):

- The place area dataset represents the areas of about 2,100 Austrian communes (source: http://geoland.at).
- We computed the built-up areas from OSM data.
- In order to assess the results, we need a reference dataset. This dataset was constructed from a coordinate database containing over 17,000 Austria localities provided by "Statistik Austria". The coordinates of those localities that have the same name as a commune and that were located within this commune were determined. 1,752 relations could be computed by this approach. The following tests are performed for those communes.

Figure 7. Test datasets around the City of Linz: place areas (green areas with red border), built-up areas (orange) and reference point (green symbols).



4.1 Evaluation without Data Preparation

The first test was done without the data preparation steps from Section 3.2. Table 1 shows the results. The computation of the place location by the centroid of the place area (i.e. the commune) results in an average distance of about 1741m to the reference point. Using the visual center instead, the average distance is approximately halved to 840m. A measure for the best possible value that the presented algorithm can theoretically achieve is the minimum distance between the largest components of the built-up area (within the place area) to the reference point. This distance is on average about 412m.

Table 1: Average distances (in meters) for Austria dataset.

	Centroid	Visual Center	Best Possible
Average Distance (m)	1741.45	840.64	412.31

4.2 Evaluation of the Removal of Holes

Table 2 shows the results when removing holes in a preprocessing step. The upper row refers to a removal of holes with an area smaller than a given absolute value. The lower row gives the results with a relative threshold in respect to the area of the whole polygon. In both cases, we can observe that larger thresholds have low impacts compared to the next smaller threshold.

Table 2: Average distances (in meters) for Austria dataset with removal of holes.

	no removal	0.01km²	0.0625km²	0.25km²	1km²
Austria dataset	840.64	835.94	822.45	823.43	822.42
	no removal	1.0%	2.5%	5.0%	10%
Austria dataset	840.64	837.12	824.08	821.54	820.77

According to Table 2, the overall impact seems to be low. This observation changes when we subdivide the dataset: the subset S10 contains the 10% quantile of small built-up areas and the subset L10 the 10% quantile of large built-up areas. Table 3 depicts the results. For subset S10, the removal of holes has no or sometimes negative impacts. The reason is obvious: most of these polygons have no holes. In contrast, subset L10 gains from the removal of small holes. For large holes, no or mixed impacts can be observed. Overall, we can conclude that the most robust approach is remove holes with low relative and low absolute thresholds; in the following, thresholds of 0.0625km² and 5% are used.

Table 3: Average distances (in meters) for subsets of the Austria dataset with removal of holes.

	no removal	0.01km²	0.0625km²	0.25km²	1km²
subset S10	742.11	742.11	747.15	747.15	747.15
subset L10	964.1	944.09	895.48	915.48	905.86
	no removal	1.0%	2.5%	5.0%	10%
subset S10	742.11	747.15	747.15	747.15	747.15
subset L10	964.1	900.48	900.48	900.48	900.48

4.3 Evaluation of the Removal of Indentations

The indentations are removed by performing two buffer operations with distance d and $-d$. Remaining holes are removed afterwards. Distance d can be determined by an absolute or by a relative value. The following relative value refers to one quarter of the perimeter of the minimum bounding box that includes the built-up area of a commune. Table 4 shows the results for both variants for the whole Austria dataset and the two subsets specified in the previous subsection.

Table 4: Average distances (in meters) for Austria dataset (whole & subsets) with removal of indentations.

	no removal	50m	100m	250m	500m
Austria dataset	840.64	808.16	795.09	799.18	916.64
subset S10	742.11	767,91	712,09	653.26	663.95
subset L10	964.1	871,34	878.6	853.57	1067,91
	no removal	0.25%	1.0%	2.5%	5.0%
Austria dataset	840.64	804.39	785.17	837.79	925.7
subset S10	742.11	742,20	728.73	684,97	631,89
subset L10	964.1	812.7	843.67	952,48	1213.01

According to these results, relative values of d do not serve both small and large polygons. Therefore, the use of an absolute distance d seems to be more suitable. Furthermore, buffer distances chosen too large have negative impacts on the results: an average distance of 1067m for the L10 subset is considerably worse than all other results. Figure 8 illustrates that buffer distances of 500m or more significantly falsify the shape of the built-up areas. For the examples in Section 4.4, a distance d of 100m is used.

Because there are no communes without built-up areas in the test dataset, the last problem addressed in Section 3.2 was not investigated.

Figure 8. Original built-up areas (dark green) and added areas after the removal of indentations (colors according to the legend) for Linz area.

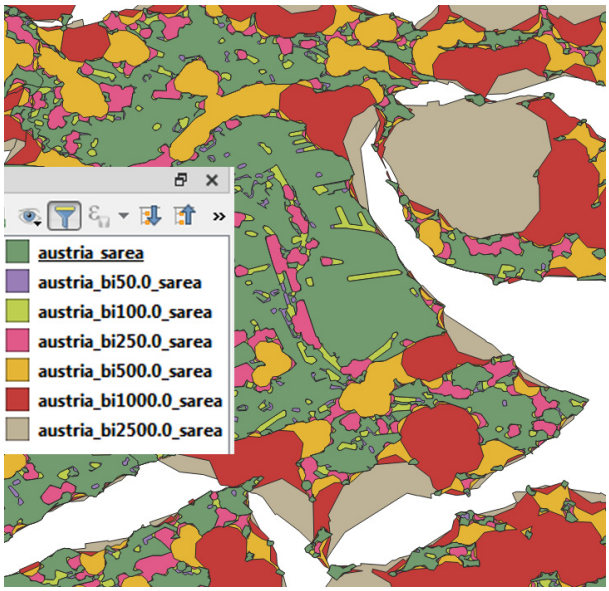
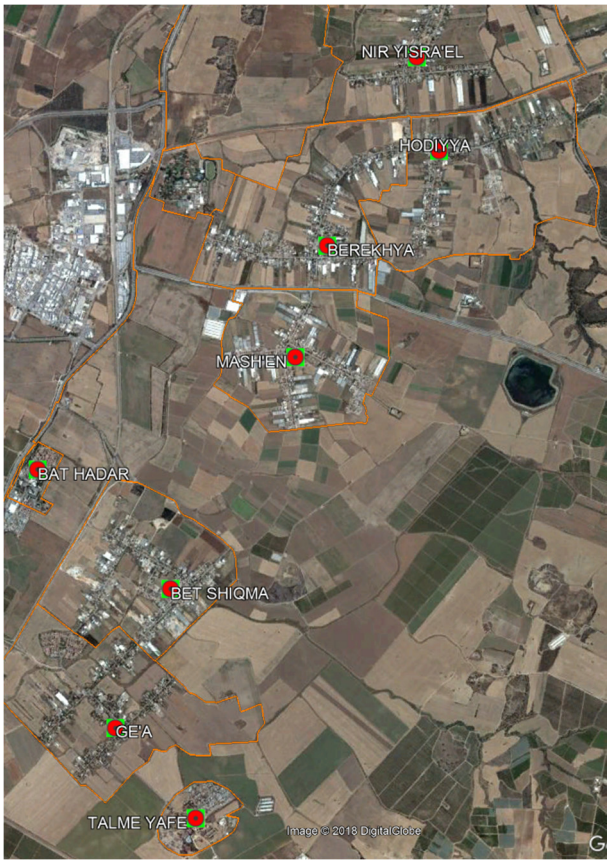


Figure 9. Part of the result of the Israel dataset (background: © 2018 DigitalGlobe)



4.4 Application

In first applications, we used the proposed method to compute locations from area datasets of Israeli and Portuguese places. In case of the Israel dataset, we were able to compute a location for 1,088 of 1,094 populated places (see also Figure 9) by using OSM built-up areas. For the Portugal dataset, OSM areas were not sufficient. Using the GUF dataset instead, we could determine locations of 16,424 of 16,953 places on the mainland of Portugal with at least 50 inhabitants (= 96.9%). Figure 10 depicts one place area and computed location.

Figure 10. Result for a village in Portugal (background: © 2018 Google).

5 Conclusions

In this paper, we presented a technique for determining point locations of places by using place areas and built-up areas. The impacts of preprocessing steps and of their parameterization on the quality of results were investigated. First applications of the method show promising results.

The evaluation is preliminary; it shall be extended by investigating further place datasets and by using GUF data in addition. The best possible value the presented algorithm can theoretically achieve for the investigated dataset (412m, see Table 1) is relatively high. In contrast, the minimum distance between any (instead of the largest) component of the built-up area to the reference point is much lower (129m). This observation indicates that the (direct and indirect) selection of the largest component for computing the visual center is too restricted. It seems that we need at least a second criterion that considers, e.g., the centering of the component.

For the case that more information is available (e.g., the population density or the position of the town hall), place areas can be reduced in size and the proposed algorithm would probably compute improved results.

A further application of the proposed algorithm is possible: If a dataset is available with place coordinates of varying accuracy (see Section 2.1), the computed intersection between place area and built-up area can be used for assessing the coordinate: If it does not lie within the calculated polygon, a mark can be assigned that the coordinate is of questionable quality.

References

Agafonkin, V. (2016) A new algorithm for finding a visual center of a polygon. https://blog.mapbox.com/a-new-algorithm-for-finding-a-visual-center-of-a-polygon-7c77e6492fbc

Asano, T. & Toussaint, G. T. (1986) Computing the geodesic center of a simple polygon. *Perspectives in Computing: Discrete Algorithms and Complexity*, Proceedings of Japan-US Joint Seminar, pp. 65-79.

Bhatti, S. S. & Tripathi, N. K. (2014) Built-up area extraction using Landsat 8 OLI imagery. *GIScience & Remote Sensing*, 51(4), pp. 445-467.

Brinkhoff, T. (2016) Open Street Map Data as Source for Built-up and Urban Areas on Global Scale. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume XLI-B4, 2016, pp. 557-564, doi:10.5194/isprs-archives-XLI-B4-557-2016

Dorn, H., Törnros, T. & Zipf, A., 2015. Quality Evaluation of VGI Using Authoritative Data – A Comparison with Land Use Data in Southern Germany. *ISPRS International Journal of Geo-Information*, 4(3), pp. 1657-1671.

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., Schwinger, M., Taubenböck, H., Müller, A. & Dech, S. (2013): Urban Footprint Processor – Fully Automated Processing Chain Generating Settlement Masks From Global Data of the TanDEM-X Mission. *IEEE Geoscience and Remote Sensing Letters*, 10 (6), 2013, 1617-1621, doi:10.1109/LGRS.2013.2272953

Estima, J. & Painho, M. (2013) Exploratory analysis of OpenStreetMap for land use classification. *Proceedings 2nd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pp. 39-46.

Garcia-Castellanos, D. & Lombardo U. (2007) Poles of Inaccessibility: A Calculation Algorithm for the Remotest Places on Earth. *Scottish Geographical Journal*, 123 (3), 227-233.

Jokar Arsanjani, J., Mooney, P., Zipf, A. & Schauss, A. (2015) Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. In: Jokar Arsanjani, J. et al. (eds.), *OpenStreetMap in GIScience*, Springer, pp. 37-58.

Miyazaki, H., Shao, X., Iwao, K. & Shibasaki, R. (2014) Development of a Global Built-Up Area Map Using ASTER Satellite Images and Existing GIS Data. In: Weng, Q. (ed.), *Global Urban Monitoring and Assessment through Earth Observation*, CRC Press, pp. 121-142.

Pollack, R., Sharir, M. & Rote, G. (1989) Computing the Geodesic Center of a Simple Polygon. *Discrete & Computational Geometry*, 4, pp. 611-626

Zhang, J., Li, P. & Wang, J. (2014) Urban Built-Up Area Extraction from Landsat TM/ETM+ Images Using Spectral Information and Multivariate Texture. *Remote Sensing*, 6(8), pp. 7339-7359.