Journal of Spatial Information Science
Number N (YYYY), pp. xx–yy

doi:10.5311/JOSIS.YYYY.II.NNN

Research Article

Bananas and tangerines spilled on streets

Martin Fleischmann

Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool, United Kingdom

Anastassia Vybornova

NEtworks, Data and Society (NERDS), Computer Science Department, IT University of Copenhagen, Denmark

Received: December 24, 2015; returned: February 25, 2016; revised: July 13, 2016; accepted: September 5, 2016.

Abstract: 1-2 sentences basic introduction into field. 2-3 sentences more detailed background. 1 sentence clearly stating the general problem being addressed by this particular study. 1 sentence summarizing the main result ("here we show"). 2-3 sentences explaining what the main result reveals/adds. 1-2 sentences to put results into more general context. Optional - if accessibility is enhanced by this: 2-3 sentences to provide broader perspective.

Keywords: street networks, blocks, urban form, shape analysis, urban morphology, urban morphometrics, routing

1 Introduction

Cities have been the object of scientific inquiry for thousands of years [22], particularly for their growth dynamics, population development, and spatial structure. Within the last 50 years, powered by the emergence of Big Data and computational power, data-driven approaches to the study of cities have gained importance. In this context, the urban street networks have proven to be a particularly useful type of data. Street networks are line-based abstractions of the street space, containing information on the connections between intersections and street segments. They are popular objects of urban analyses due to their explanatory power and simplicity (digitizing a street network is much easier than digitizing buildings). At the same time, feasibility of studies that take street networks as a

Figure 1: [A close-up look at [city, location]. The black lines are network edges (street segments). Grey polygons are correctly identified urban blocks; the red polygon is a face artefact that should not be identified as urban block.]

point of departure has greatly increased with open source GIS data becoming available on platforms like OpenStreetMap [2]. Street networks have a wide range of applications, from transportation network design [9], assessment of urban sprawl [3] or evolution and distribution of street patterns [4,5] to the classification of urban form [1,10].

Street network data sets vary greatly in their level of detail and data quality [CITE]. Whether a street network data set is fit for purpose depends to a large extent on the specific application. For example, traffic routing applications require adequately represented directionality of edges (street segments) [CITE], while urban morphology studies are based on the shape of urban blocks (i.e. the polygons between the edges), and thus do not require the edges to be directed [7]. A frequent street network data processing challenge, common to many applications, is to reduce granularity of detail without loss of relevant information and introduction of new imprecisions. In many cases, deciding which information to keep and which to aggregate is easy for a human, but challenging for an algorithm (see Figure 1). This is true both for studies that are concerned with the street network itself, and for studies that look at the polygons enclosed by the street network. Further complexity is added by the fact that the level of detail of the input network might vary greatly depending on the data source.

In short, working with street networks entails not only valuable insights, but also several much-lamented, yet unresolved methodological challenges. In this article, we aim to tackle one of these challenges, namely the detection and removal of what we call *street network face artefacts*, or short: *face artefacts*, as explained in detail in the section below.

The paper is organized as follows: In the next sections, we expand on the problem and the key questions, followed by a review of literature and terminology. We then introduce our methodology, proposing a cheap computational heuristic which allows the automatized identification of face artefacts based on a shape index. We then briefly describe our workflow for face artefact detection, which we apply to XXX cities across the globe. Lastly, we present our results and conclude by discussing implications and outlining potential further steps.

Problem description

Each geospatially encoded street network comes with a certain level of granularity and a focus on specific elements of street space. While all attempt to capture primarily connectivity, the resulting graphs can vastly differ. In graph theory, the polygons enclosed by the network edges in a planar space are called *(graph) faces* [CITE]. A detailed look at these face polygons for a given street network often reveals artifacts of transport-focused geometry, as illustrated in Figure 1.

Unlike its grey-colored neighbors, the face polygon colored in red is not an urban block enclosed by streets; rather, it appears in the network due to the representation of a bidirectional street as two separate edges, one in each direction of traffic flow. This way of network representation is suitable for car-based routing, but much less suitable for other applications. If our goal is to generate polygons that are representative of urban blocks and inversely to ensure that the graph is representing the morphological network rather than the transportation network, we can call such face polygons polygons "face artifacts", as they occur only as a result of the data preparation model not suited for the purpose. Face artifacts pose a twofold problem. First, for studies concerned with urban form, they introduce a false signal into the distribution of urban shapes and distort the actual shape of their neighboring polygons. Second, for studies concerned with the properties and pattern of the street network, face artefacts introduce superfluous network edges, thus distorting all network metrics based on node degree and/or shortest path computations. A further aggravating factor is that the extent to which face artifacts distort results depends on the analysis conducted, and cannot be quantified without prior identification of such polygons. Thus, no matter whether one is interested in the urban street network or in urban shapes enclosed by the network, the face artifact should be removed as part of data preprocessing, and replaced by a single network edge. Human manual identification of face artefacts would be unambiguous, but prohibitively costly, not scalable and not entirely reproducible. Although this issue has been pointed out by many authors (see section below), a fully automatized approach to the removal of face artefacts is, to our knowledge, still non-existent. We therefore pose the following research question:

How can face artifacts in an urban street network be computationally identified?

In this article, we propose a method to answer this question and test the proposed method's universality.

Literature review

While face artefacts are a commonly known problem in the research community, there is a lack of coherent terminology for the phenomenon. Previous studies have referred to the same issue in widely varying terms, which makes it substantially more difficult to conduct a comprehensive literature review on the topic. Hereby, we apologize for any involuntary omissions of previous work on face artefacts, and simultaneously wish to contribute to a future homogenization of the terminology.

Few studies that explicitly tackle the "face artifact" issue could be identified.

Li et al. [14] point out the difficulties of extracting multilane roads from OpenStreetMap (OSM) that arise from each lane being represented as a separate linestrings. The authors propose a method to identify and merge so-called "multilane polygons", i.e. adjacent polygons covering a single street area that result from mapping of multiple street lanes, through a SVM (support vector machine) machine learning algorithm that uses five shape parameters as input. While this method does succeed in identifying face artefacts at multilane roads, it is only reproducible by users with advanced machine learning skills; furthermore, the method requires input of manually classified training data, which adds a substantial amount of effort.

In their study on feature matching between OSM and reference data, Fan et al. [8] identify face artifacts, which they call "non-urban block polygons", as a data preprocessing issue. The authors use the SVM approach developed by Li et al. [14], as mentioned above, to identify face artefacts; they point out that the approach fails for smaller face artefacts at traffic junctions.

Sanzana et al. [19] elaborate on the process of deriving hydrological response units from drainage networks and find that error correction is needed for so called "bad-shaped polygons". One of the subcategories of bad-shaped polygons, as classified by the authors, is "sliver polygons", formed mainly by roads and footpaths.

Grippa et al. [12] classify polygons derived from OpenStreetMap street network data into "urban blocks" and "sliver polygons" and present a semi-automated workflow, partially in PostGIS, for sliver polygon removal from the data set. Ludwig et al. [16] take up this approach within the context of land use classification and additionally filter polygons based on a size threshold.

Vybornova et al. [23] refer to the network pattern that creates "face artifact" as "parallel edges" and present a network shortest path-based approach for their identification, but no solution to effectively remove these from the network.

Related, but not identical to "face artifact" is the problem of formalization of street space. In this regard, Peponis et al. [17] conduct an analysis of urban spatial profiles and point out that their input data includes street center lines, but lacks street widths, hence street surfaces are merged into urban blocks.

In a methodologically different approach, Hermosilla et al. [13] develop a method to derive so-called urban block related street areas (abbreviated by the authors as "UBRSA"), defined as the street area surrounding an urban block. This method, however, requires urban block boundaries as an input.

Lastly, a recent study by Shpuza [20] describes elongated urban blocks that are delimited either by a street or another type of obstacle (e.g. a waterbody), and that contain no buildings, as "edge blocks". Edge blocks can be identified as outliers in a so-called shape matrix based on two geometrical parameters, relative distance and directional fragmentation. However, edge blocks represent actual urban blocks rather than scattered parts of the street space.

A side note on terminology

Some recent studies describe the "face artifact" phenomenon as sliver polygons [12,16,19].

However, "face artifact" arise as consequence of a context-dependent redundancy of mapped line features, while sliver polygons stem from mismatching boundaries in vector overlays of polygon features [6, 11, 18]. The other suggestions available in literature are not suitable either. "Multilane polygons" [14] or "parallel edges" [23] do not reflect other transportation geometries causing the issue (e.g. complex intersections) while "bad-shaped polygons" use a relatively vague term "bad" that does not indicate the actual issue. In addition, "face artifact" are, in line with our problem definition, confined to the specific context of urban street networks. Therefore, in spite of some degree of geometric similarity between the two, we refrain from applying the term "sliver polygon" in the "face polygon" context and rather build on more generic terminology derived from graph theory.

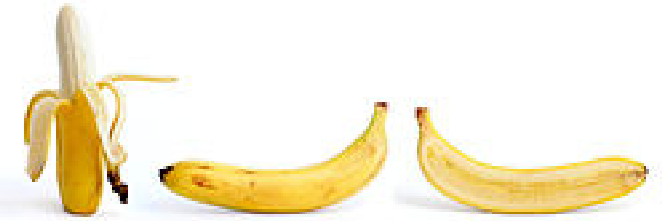


Figure 2: [Show shape index v. area plot for different continents (?) or for different FUA(s) to show the banana shape in the distribution]

2 Method

Make a strong case for the fact that our method is very simple (it is not a machine learning algorithm); computationally cheap; AND manages to capture BOTH elongated bananas and intersection bananas with ONE stroke and ONE index, which is possible thanks to the CHARACTERISTIC PATTERNS in urban street networks. As demonstrated in the scatter plot 2...

Method section on using shape indeces

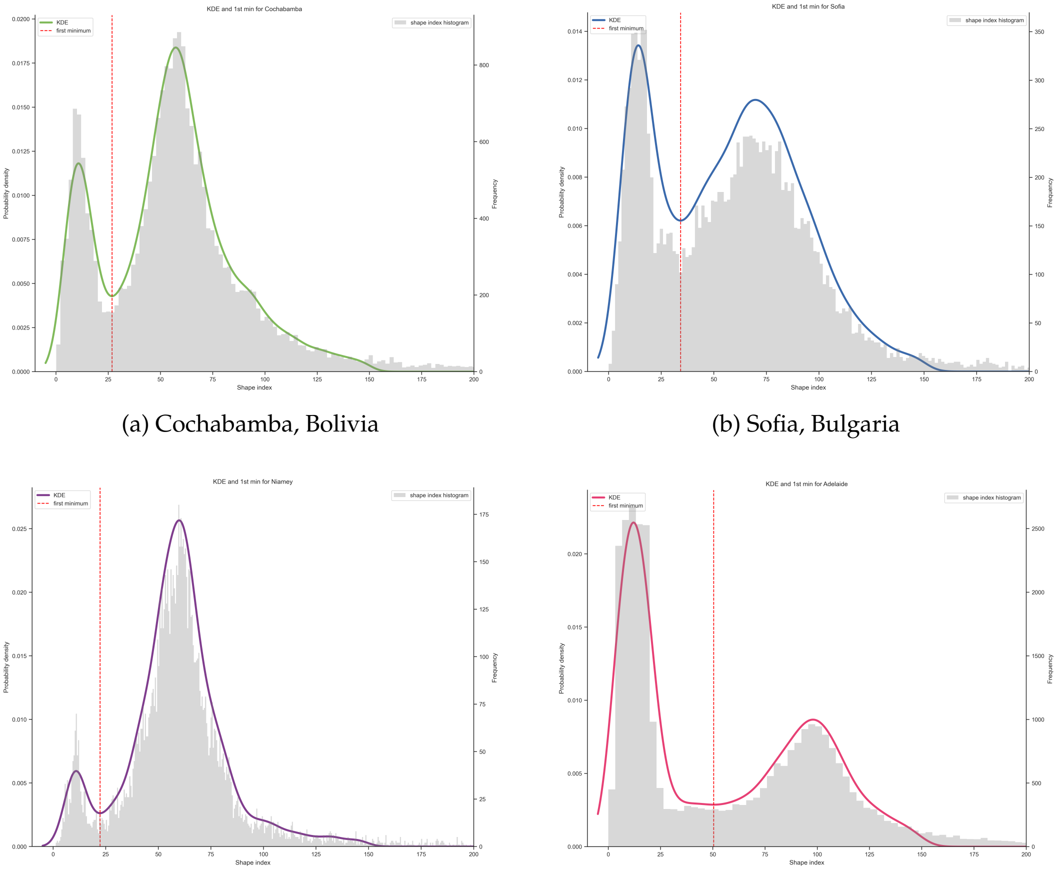
mention Sanzana [19], Louf [15], ... etc.

*Sanzana et al. propose an algorithm for identification and elimination of "bad-shaped polygons", incl. streets/roads/footpaths. Explain that while we have a comparable approach, we are concerned with *cities* while they are concerned with *hydrological models*; and since cities express some certain regularities we can make use of that*

Method section on finding the minimum and using it as a threshold (put part of this in results maybe?)

Many, though not all, shape index frequency distributions for the analyzed FUAs reveal a common feature of two prominent peaks (see Figure **??**). Through visual analysis, we find that these peaks represent two different types of polygons. Most of the polygons from the first (leftmost) peak can be attributed to "bananas" in the street network, whereas most of the polygons from the second (rightmost) peak represent true urban blocks. Therefore, for FUAs that show a pronounced two-peak pattern in their shape index frequency distribution, the minimum *between* the two peaks can be used as shape index threshold: polygons with a shape index below the threshold will most likely be "bananas"; polygons with a shape index above the threshold will most likely be true urban blocks. To derive the minimum, we approximate the shape index frequency distribution with a Gaussian kernel density estimation. For bandwidth selection, we use the parametric Silverman method [21] and find that it gives satisfactory results; non-parametric bandwidth selection methods might be a subject for future work (see section 4).

Next, comparing the positions of peak 1, peak 2, and shape index threshold for different FUAs with pronounced two-peak patterns, we find that maxima positions vary to a con-



(a) Cochabamba, Bolivia

(b) Sofia, Bulgaria

(c) Niamey, Niger

(d) Adelaide, Australia

Figure 3: Shape index frequency distributions of face polygons for different continents.

Figure 4: [Demonstrating that it works: Show example of all bananas in 1 city (big), plus a zoom-in of a particularly banan-y area, and the corresponding shape index distribution plus threshold (small), to demonstrate that the first peak corresponds to (mostly) bananas]

siderably greater extent than minima positions. In other words, shape index thresholds for "banana" identification from morphologically different FUAs lie within a relatively narrow range. We therefore hypothesize that applying a shape index threshold within the range identified from FUAs whose polygons follow a two-peak distribution will allow the identification of "bananas" polygons even for those FUAs whose distributions do not show a two-peak pattern. Applying the ... [TBD: lower boundary/median/average/higher boundary] of the empirically derived shape index threshold range to the rest of FUAs indeed reveals "bananas´´ polygons in all/most FUAs (see Figure 4).

3 Results

- area vs shape plots - use all cases together and show multiple shape indices - Reock as an optimal index (?) [I think it will be the optimal one but we need to verify that] - 1-dimensional index formula (if we use Reock it is the one from the banana notebook) - shape-index plots with cut-off values - plots based on geographical location - distributions, Reock-area scatters - describe the differences - formalise the detection workflow

4 Discussion

How could this be used?

how to move forward? (sneak preview of google summer of code) - the simplification problem can be seen as a problem of the elimination of banana

incorporate further data (ideas: directionality; street names; angles; land use; ...) use network formalism: on dual approach (intersections = edges): jiang 2004, yang 2022, rosvall/sneppen; barthelemy paper on shortest path shape

end with a call to action & 'towards open urban data science'

include in future work:

- analyze other regularities in distribution
- !! once banana has been found: how to replace it?
- non-parametric bandwidth selection
- using data on land use of potential "bananas" to identify whether they are urban blocks or not - would be great IF data was there (as discussed by Fan et al. [8])

Data and code availability

The fully documented workflow, all input data and all results are made available in open source format on GitHub: github.com/martinfleis/bananas.

Acknowledgments

To be added. Remember to include ESRC/ATI funding covering initial experiments.

References

[1] ARALDI, A., AND FUSCO, G. From the street to the metropolitan region: Pedestrian perspective in urban fabric analysis:. *Environment and Planning B: Urban Analytics and City Science 46*, 7 (Aug. 2019), 1243–1263. 10.1177/2399808319832612. tex.ids: araldi2019a.

[2] ARCAUTE, E., AND RAMASCO, J. J. Some recent advances in urban system science: models and data, Dec. 2021. Number: arXiv:2110.15865 arXiv:2110.15865 [physics].

[3] BARRINGTON-LEIGH, C., AND MILLARD-BALL, A. Global trends toward urban street-network sprawl. *Proceedings of the National Academy of Sciences 117*, 4 (Jan. 2020), 1941–1950. 10.1073/pnas.1905232116.

[4] BOEING, G. A multi-scale analysis of 27,000 urban street networks: Every US city, town, urbanized area, and Zillow neighborhood. *Environment and Planning B: Urban Analytics and City Science 219*, 4 (Jan. 2018), 239980831878459. 10.1177/2399808318784595.

[5] BOEING, G. Off the grid… and back again? The recent evolution of american street network planning and design. *Journal of the American Planning Association* (2020), 1–15. 10.1080/01944363.2020.1819382. Publisher: Taylor & Francis.

[6] DELAFONTAINE, M., NOLF, G., VAN DE WEGHE, N., ANTROP, M., AND DE MAEYER, P. Assessment of sliver polygons in geographical vector data. *International Journal of Geographical Information Science 23*, 6 (June 2009), 719–735. 10.1080/13658810701694838. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13658810701694838.

[7] DIBBLE, J., PRELORENDJOS, A., ROMICE, O., ZANELLA, M., STRANO, E., PAGEL, M., AND PORTA, S. On the origin of spaces: Morphometric foundations of urban form evolution. *Environment and Planning B: Urban Analytics and City Science 46*, 4 (May 2019), 707–730. 10.1177/2399808317725075.

[8] FAN, H., YANG, B., ZIPF, A., AND ROUSELL, A. A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data. *International Journal of Geographical Information Science 30*, 4 (Apr. 2016), 748–764. 10.1080/13658816.2015.1100732.

[9] FARAHANI, R. Z., MIANDOABCHI, E., SZETO, W. Y., AND RASHIDI, H. A review of urban transportation network design problems. *European Journal of Operational Research 229*, 2 (Sept. 2013), 281–302. 10.1016/j.ejor.2013.01.001.

[10] FLEISCHMANN, M., FELICIOTTI, A., ROMICE, O., AND PORTA, S. Methodological foundation of a numerical taxonomy of urban form. *Environment and Planning B: Urban Analytics and City Science* (Dec. 2021), 239980832110598. 10.1177/23998083211059835.

[11] GOODCHILD, M. F. Statistical aspects of the polygon overlay problem. *Harvard papers on geographic information systems* (1978). Publisher: Addison-Wesley.

[12] GRIPPA, T., GEORGANOS, S., ZAROUGUI, S., BOGNOUNOU, P., DIBOULO, E., FORGET, Y., LENNERT, M., VANHUYSSE, S., MBOGA, N., AND WOLFF, E. Mapping Urban Land Use at Street Block Level Using OpenStreetMap, Remote Sensing Data, and Spatial Metrics. *ISPRS International Journal of Geo-Information 7*, 7 (July 2018), 246. 10.3390/ijgi7070246. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[13] HERMOSILLA, T., PALOMAR-VAZQUEZ, J., BALAGUER-BESER, A., BALSA-BARREIRO, J., AND RUIZ, L. A. Using street based metrics to characterize urban typologies. *Computers, Environment and Urban Systems 44* (Mar. 2014), 68–79. 10.1016/j.compenvurbsys.2013.12.002.

[14] LI, Q., FAN, H., LUAN, X., YANG, B., AND LIU, L. Polygon-based approach for extracting multilane roads from OpenStreetMap urban road networks. *International Journal of Geographical Information Science 28*, 11 (Nov. 2014), 2200–2219. 10.1080/13658816.2014.915401.

[15] LOUF, R., AND BARTHELEMY, M. A typology of street patterns. *Journal of The Royal Society Interface 11*, 101 (Dec. 2014), 20140924. 10.1098/rsif.2014.0924. Publisher: Royal Society.

[16] LUDWIG, C., HECHT, R., LAUTENBACH, S., SCHORCHT, M., AND ZIPF, A. Mapping Public Urban Green Spaces Based on OpenStreetMap and Sentinel-2 Imagery Using Belief Functions. *ISPRS International Journal of Geo-Information 10*, 4 (Apr. 2021), 251. 10.3390/ijgi10040251.

[17] PEPONIS, J., ALLEN, D., HAYNIE, D., SCOPPA, M., AND ZHANG, Z. Measuring the Configuration of Street Networks: The Spatial Profiles of 118 Urban Areas in the 12 Most Populated Metropolitan Regions in the US. In *Proceedings of the 6th International Space Syntax Symposium* (Istanbul, Turkey, 2007), Istanbul Technical University, p. 17.

[18] RYBACZUK, K. Using information based rules for sliver polygon removal in GISs. In *Geographic Information Systems, Spatial Modelling and Policy Evaluation*, M. M. Fischer and P. Nijkamp, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993, pp. 85–102.

[19] SANZANA, P., GIRONÁS, J., BRAUD, I., HITSCHFELD, N., BRANGER, F., RODRIGUEZ, F., FUAMBA, M., ROMERO, J., VARGAS, X., MUÑOZ, J. F., VICUÑA, S., AND MEJÍA, A. Decomposition of 2D polygons and its effect in hydrological models. *Journal of Hydroinformatics 21*, 1 (Sept. 2018), 104–122. 10.2166/hydro.2018.031.

[20] Shpuza, E. The shape and size of urban blocks. *Environment and Planning B: Urban Analytics and City Science* (May 2022), 239980832210987. 10.1177/23998083221098744.

[21] Silverman, B. W. Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society: Series B (Methodological) 43*, 1 (1981), 97–99. 10.1111/j.2517-6161.1981.tb01155.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1981.tb01155.x.

[22] Vitruvius. *Vitruvius: 'Ten Books on Architecture'*. Cambridge University Press, 1999. 10.1017/CBO9780511840951.

[23] Vybornova, A., Cunha, T., Gühnemann, A., and Szell, M. Automated Detection of Missing Links in Bicycle Networks. *Geographical Analysis n/a*, n/a (2022). 10.1111/gean.12324. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gean.12324.