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

Many studies within urban science use the street network of a city as primary input. Recent examples are numerous and cover a wide range of applications: from transportation network design [5] to the classification of urban morphology [6]. The feasibility and broad applicability of quantitative urban science studies has greatly increased with open source GIS data becoming available on platforms like OpenStreetMap [1].

However, even when a data set of sufficient quality is available, the conversion of geospatial features into a street network, sometimes referred to as "network simplification", still poses several unresolved methodological questions to the research community. The challenge, in a nutshell, is to reduce granularity of detail without loss of relevant information. In many cases, deciding which information to keep and which to aggregate is easy

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 "banana" that should not be identified as urban block.]

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 requirements towards the input network vary greatly depending on the use case. For example, traffic routing applications require adequately represented directionality of edges (street segments), while urban morphology studies are based on the shape of urban blocks, i.e., the polygons between the edges [3].

In this article, we work towards resolving one specific issue within the features-to-network/features-to-blocks conversion process, which arises from a transportation-focused mapping of urban space and which we call "bananas". The paper is organized as follows: In the next sections, we define the problem and formulate our research question, followed by a literature and terminology review. We then briefly describe the methodology and our workflow that we apply to X cities across the globe, and present an overview of our results. The fully documented workflow, all input data and all results are made available in open source format on GitHub: github.com/martinfleis/bananas. We conclude with a discussion of implications and potential further steps.

Problem description

Each geospatially encoded street network comes with a certain level of granularity. A detailed look at the polygons enclosed by the edges (street segments) of a given network often reveals artifacts of transport-focused geometry. An illustration is found in Figure 1.

Unlike its grey-colored neighbors, the 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. The "banana" in Figure 1 poses a twofold problem. First, for studies concerned with urban form, it introduces a false signal into the distribution of urban shapes and distorts the actual shape of its neighboring polygons. Second, for any study that is concerned with the properties and pattern of the street network rather than with routing, the "banana" introduces a superfluous network edge, and it does so in a frequently inconsistent matter – not for all, but only some bidirectional and/or multilane streets. The extent to which "banana" artifacts distort results depends on the analysis conducted, and cannot be quantified without prior "banana" identification. Thus, no matter whether one is interested in the urban street network or in shapes enclosed by the network, the "banana" should be removed as part of data preprocessing, and replaced by a single network edge. Human manual processing would be unambiguous, but prohibitively costly. We therefore pose the following research question:

How can "bananas" in an urban street network be computationally identified?

The rest of the paper aims at answering this question and is organized as follows: first, we conduct a brief literature review and outline terminological "banana" issues. Then, we propose a simple computational heuristic which allows the identification of "bananas" based on the "banana index". Next, we apply the proposed method to X functional urban areas (FUAs) across the globe and present the obtained results. We conclude with a discussion and suggestions for future work.

Literature review

First off, a disclaimer: One of the present challenges of "bananas" is the lack of a coherent terminology of the problem description (see section below), which makes it substantially more difficult to conduct a comprehensive literature review on the topic. We therefore apologize for any involuntary omissions of previous work on "bananas", and hope to contribute to a future homogenization of the problem discussion.

Few studies that explicitly tackle the "bananas" issue could be identified. Li et al. [10] identify 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. Fan et al. [4] identify "bananas" as a data preprocessing issue for their feature matching workflow, and point out that polygons derived from the street network can be classified either as urban block polygons or as road area polygons; the latter are then identified and removed based on the SVM approach developed by Li et al. [10], as reviewed above. Sanzana et al. [15] 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. [8] classify polygons derived from OpenStreetMap street network data into "urban blocks" and "sliver polygons" and present a semiautomated workflow, partially in PostGIS, for sliver polygon removal from the data set. Ludwig et al. [12] take up this approach within the context of land use classification and additionally filter polygons based on a size threshold. Vybornova et al. [18] refer to the network pattern that creates "bananas" 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 "bananas" is the problem of formalization of street space. In this regard, Peponis et al. [13] 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. [9] 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 input.

Lastly, a recent study by Shpuza [16] 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.

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

Terminology

Some recent studies describe the "banana" phenomenon as sliver polygons [8,12,15]. However, "bananas" 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 [2,7,14]. In addition, "bananas" 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 "bananas" context.

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 [15], Louf [11], ... 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 3). 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

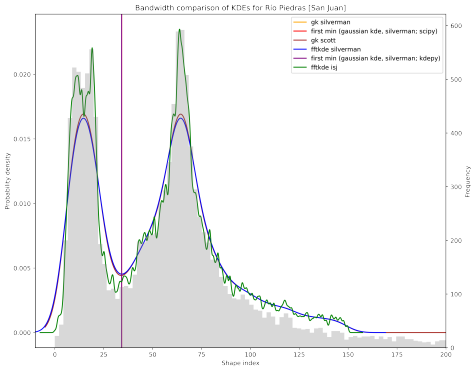

Figure 3: [Above: just a placeholder. Insert here: shape index frequency distribution(s) for FUA X. - show the two-peak pattern]

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 [17] 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 considerably 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

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]

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. [4])

Acknowledgments

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

References

[1] 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].

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.