"The Great Blackbury Pie" ~ or ~ Focal Area Bias in Geographically Weighted Analysis

"The Great Blackbury Pie" ~ or ~ Focal Area Bias in Geographically Weighted Analysis

Huck, J.J.*[1], Dennis M.[1] and Labib S.M.[2]

[1] MCGIS, Department of Geography, The University of Manchester, UK
[2] Department of Human Geography and Spatial Planning, Utrecht University, Netherlands

GISRUK 2024

Summary

Focal and geographically weighted analyses are commonplace in GIS applications across many fields and disciplines. However, where such analyses are based on 'dense' datasets (e.g., a raster surface), they can suffer from an unintended bias towards the periphery of the focal zone (neighbourhood), which (counterintuitively) is exacerbated by the use of distance weighting functions. This paper serves to characterise this problem, which we call *focal area bias (FAB)*, present a proposed correction, and point to extensive simulation-based analysis, which demonstrates both the impact that this effect can have on analyses and the efficacy of our proposed solution.

KEYWORDS: Focal Analysis; Neighbourhood Analysis; Exposure Assessment; Spatial Ecology; Spatial Epidemiology

1. Introduction

The Great Blackbury Pie is a short story written in 1970 by Terry Pratchett, in which an order for one hundred pies of one foot diameter is accidentally made for a one pie of one hundred foot diameter instead, leading to calamity (Pratchett, 2023). The comedy in the story is derived from the nonlinear relationship between the area and diameter (or radius) of a circle – a situation that can also lead to calamity in spatial analysis. This paper will explore this issue, which we name *Focal Area Bias* (FAB), and seek to both characterise and resolve it through the production of a corrective function.

1.1. Focal Area Bias

Focal analyses in GIS describe a widely used set of techniques intended to characterise a feature or location based on the properties of its neighbourhood, normally through the aggregation of surrounding data using summary statistics (e.g., mean pollution value, or proportion of a given landcover class). They are distinguished from zonal analyses, which seek to characterise the zone itself. Focal approaches are widely applied in studies of environmental exposure (e.g., Labib et al., 2020), ecology (e.g., Janzén et al., 2023), landscape genetics (e.g., Vantaux et al., 2021), and urban studies (e.g., Huck et al., 2023), amongst others. Where those approaches include the use of a distance-weighting function, they are typically called geographically weighted analyses (e.g., Comber et al., 2016).

The goal of characterising a location based on its neighbourhood (as opposed to the neighbourhood zone itself) has implications for the analysis when the underlying data are *dense* (e.g., a raster surface or point grid). This is because the application of a summary statistic (e.g., mean, count or proportion) makes the implicit assumption that all locations within the focal zone are equally important, which is at odds with this goal, as is illustrated by the cumulative distance functions in **Figure 1**. Here, the steepest parts of the curve reveal the distances from the focal zone that have the greatest impact on the analysis under the above assumption (blue), and one in which each *distance* from the focal zone is

* jonathan.huck@manchester.ac.uk

equally important (in green). The fundamental issue here is simply that the nonlinear relationship between the radius and area of the zone means that there are more locations that are further away from the focal point than those that are closer, thus biasing the analysis towards data that are furthest away from the focal location (as can be seen in the blue curve).

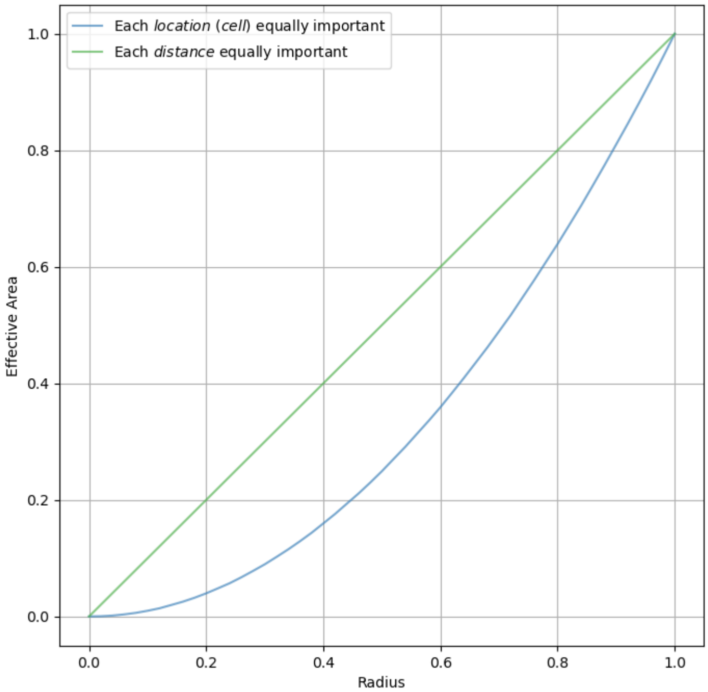

Figure 1. Cumulative distance functions, in which a steeper curve denotes more influence over the analysis at a given distance from the focal location.

The implications of this condition are illustrated with a rudimentary example in **Figure 2**, which shows an *urban-ness* calculation (after Moll et al., 2019) that simply comprises proportion of urban cells within the focal zone. This illustrates how FAB can lead to counterintuitive results, whereby the focal location in the town centre is determined to be slightly *less* urban than farmland several kilometres away (0.22 and 0.29 respectively). Clearly this finding is counterintuitive, and a similar result could be achieved with features at a range of scales (e.g., a park surrounded by an urban area, a city surrounded by ‘green belt’ etc.).

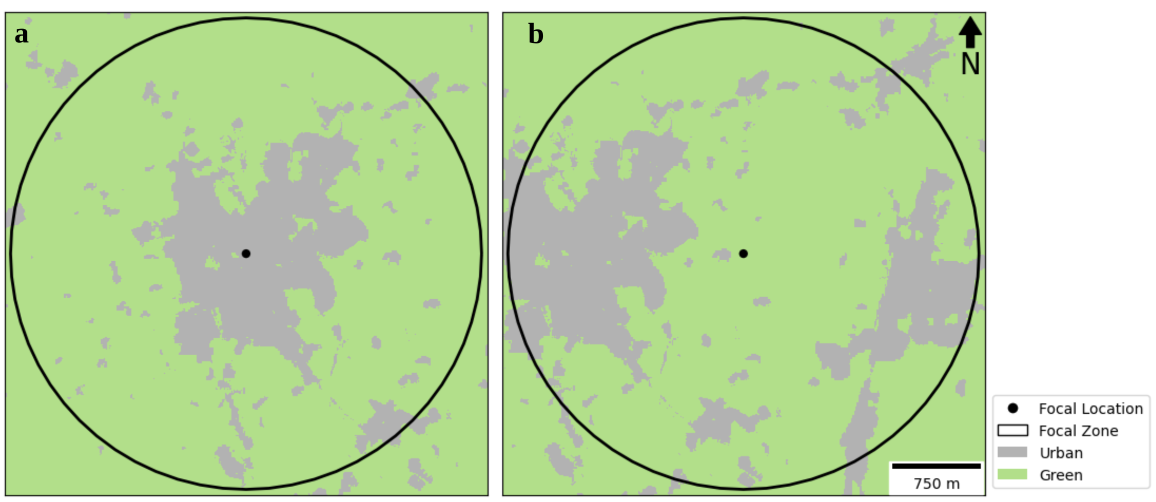

Figure 2. Two focal analyses of urban-ness**:** **a**. the town centre of Barnoldswick, Lancashire (0.22) and **b.** farmland 2km to the East of that location (0.29). Clearly, location **a** should be characterised as more urban than **b**, but the results show the opposite due to FAB.

1.2. Geographically Weighted Approaches

Though the above condition has not previously been examined in the literature, it is likely to be intuitively understood by many researchers. A frequent response is therefore to apply a distance weighting function, such as inverse distance weighting (IDW) (see Gollini et al., 2013 for a list of further examples). However, these weighting functions are also affected by the FAB, meaning that they do not give the desired effect, as shown in the cumulative distance functions in **Figure 3**. Here, the area of greatest influence (the steepest part of the curve) is moved by each function, but never to the focal location. This is due to the interaction between the weighting function and FAB and can frequently lead to further counterintuitive results whereby the weighting function *exacerbates* (rather than reduces) the bias towards other parts of the focal zone. We believe that this interaction is not widely understood by researchers, and so could negatively impact on the outcomes of focal analyses of dense datasets.

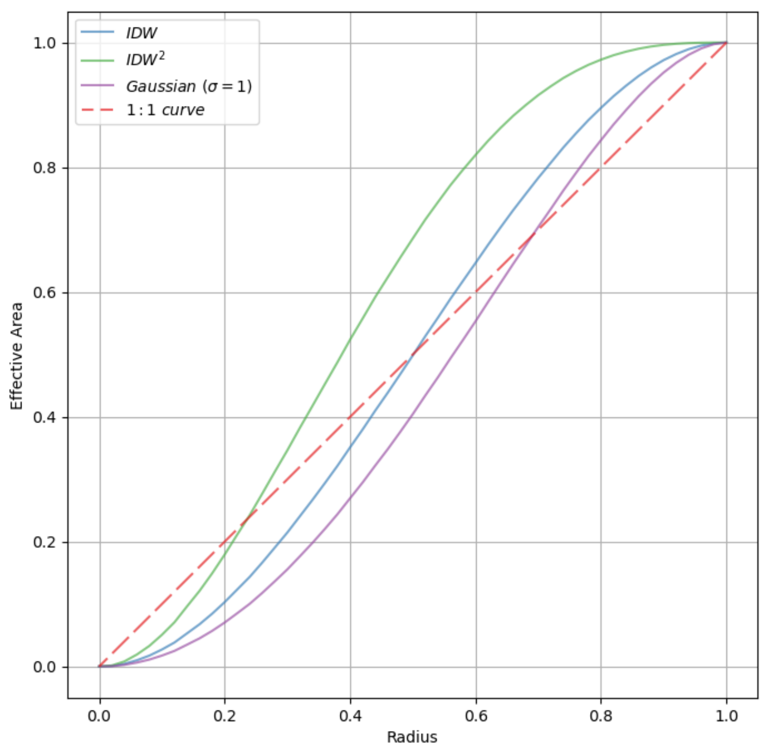

Figure 3. Cumulative distance functions showing the effect of some common weighting functions on focal analysis. A steeper curve denotes more influence over the analysis at a given distance from the focal location.

2. Correcting Focal Area Bias

FAB can be corrected using the **Equation 2**, which should be applied to each data point in a focal zone (e.g., each cell in a raster) prior to the calculation of a summary statistic.

$$c_i = \frac{a_r \, d_i}{a_d \, r} \tag{1}$$

Where: c_i is the correction for a feature at location i, d_i is the distance from the focal location to location i, r is the radius of the focal zone, a_d is the area of a buffer around the focal feature with radius d_i *and* a_r is the area of the focal zone (i.e., a buffer of radius r). In the case of focal zones constructed using regular shaped buffers where the relationship between radius area is known, such as a circular buffer around a point, then the corrective function can be approximated as per **Equation 2**, which is significantly less computationally intensive to implement (but can underperform where the focal zone is small relative to the resolution of the underlying data).

$$c_i = \frac{r}{d_i} \tag{2}$$

In either case, the correction is applied to the dataset prior to aggregation simply by multiplying each value in the focal zone (e.g., each cell in a raster surface) by c_i. The result of the corrective equation is illustrated in **Figure 4**, which clearly shows the corrective effect on an un-weighted analysis (**a**) and a selection of geographically weighted analyses (**b**).

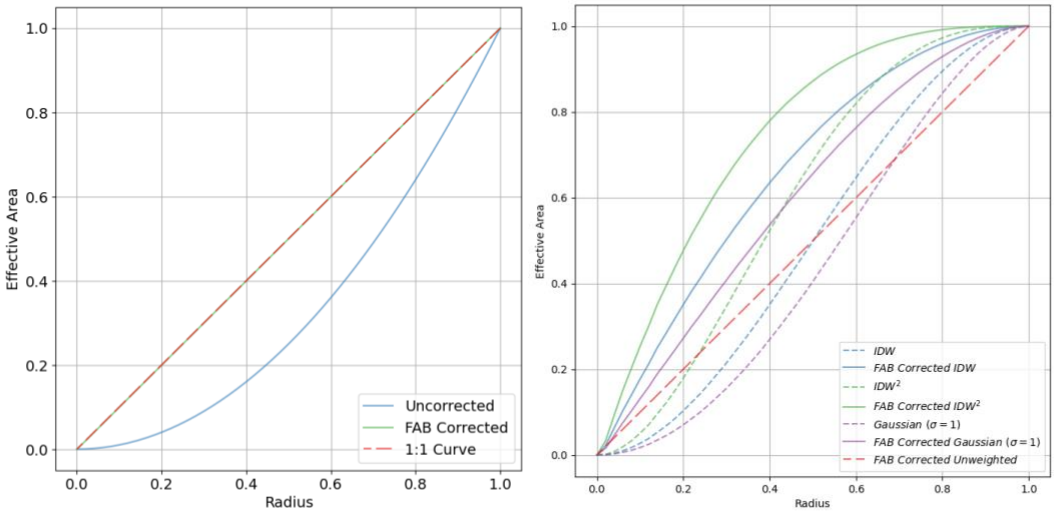

Figure 4. Cumulative distance curves showing the impact of the FAB correction on an un-weighted analysis (left), and a selection of geographically weighted focal analyses (right).

3. Conclusion

This abstract has described the problem of Focal Area Bias and the impact that it has on a range of focal and geographically weighted analyses, as well as presenting corrective functions that remove the issue. An oral presentation will include the results of a comprehensive (already completed) set of simulation-based analyses that are beyond the scope of this abstract. These analyses use both artificial and 'real' datasets on greenspace availability in Greater Manchester to demonstrate the efficacy of our proposed correction under several different scenarios, as well as characterise the magnitude of impact of FAB on the results of focal analyses.

We propose that FAB is a key methodological issue, with the potential to have substantive negative impacts on the results of focal analyses. The authors therefore recommend that FAB correction should be applied in most instances of focal analysis, particularly where geographical weighting functions are used.

4. Acknowledgements

Our thanks go to Kirsty Watkinson, Gareth Clay and Matthew Tomkins for their contributions to discussions around this research. This research was supported by the Natural Environment Research Council (grant number NE/W003120/1).

References

Comber, A., Fonte, C., Foody, G., Fritz, S., Harris, P., Olteanu-Raimond, A.-M., & See, L. (2016). Geographically weighted evidence combination approaches for combining discordant and inconsistent volunteered geographical information. *Geoinformatica*, *20*, 503-527.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2013). GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. *arXiv preprint arXiv:1306.0413*.

Huck, J., Whyatt, D., Davies, G., Dixon, J., Sturgeon, B., Hocking, B., Tredoux, C., Jarman, N., & Bryan, D. (2023). Fuzzy Bayesian inference for mapping vague and place-based regions: a case study of sectarian territory. *International Journal of Geographical Information Science*.

Janzén, T., Hammer, M., Petersson, M., & Dinnétz, P. (2023). Factors responsible for Ixodes ricinus presence and abundance across a natural-urban gradient. *PloS one*, *18*(5), e0285841.

Labib, S. M., Lindley, S., & Huck, J. J. (2020). Scale effects in remotely sensed greenspace metrics and how to mitigate them for environmental health exposure assessment. *Computers, Environment and Urban Systems*, *82*, 101501. https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2020.101501

Moll, R. J., Cepek, J. D., Lorch, P. D., Dennis, P. M., Tans, E., Robison, T., Millspaugh, J. J., & Montgomery, R. A. (2019). What does urbanization actually mean? A framework for urban metrics in wildlife research. *Journal of Applied Ecology*, *56*(5), 1289-1300.

Pratchett, T. (2023). *A Stroke of the Pen: the lost stories*. Doubleday.

Vantaux, A., Riehle, M. M., Piv, E., Farley, E. J., Chy, S., Kim, S., Corbett, A. G., Fehrman, R. L., Pepey, A., & Eiglmeier, K. (2021). Anopheles ecology, genetics and malaria transmission in northern Cambodia. *Scientific reports*, *11*(1), 1-17.

Reproducibility

All of the code used in the production of these figures, including a reference Python implementation of a Python class for FAB, are available here.

Biographies

Jonny Huck is Senior Lecturer in Geographical Information Science at the University of Manchester and Chair of the GISRUK National Steering Committee. He is interested in the creation of methods to understand uncertainty and error in GIS analysis, which he applies to a range of environmental, urban, and global health applications.

Matt Dennis is Senior Lecturer in Geographical Information Science at the University of Manchester. He does research into spatial-ecological methods and applies them to problems in the areas of nature recovery and social-ecological systems.

SM Labib is an Assistant Professor of Data Science & Health, at Utrecht University and a visiting research associate at the Public Health Modelling Group, in the MRC Epidemiology Unit, University of Cambridge. His research interests in spatial data science, GIS, and their applications in environmental epidemiology, and urban health.