AGILE: GIScience Series, 5, 21, 2024. https://doi.org/10.5194/agile-giss-5-21-2024
Proceedings of the 27th AGILE Conference on Geographic Information Science, 4–7 Sept. 2024.
Editors: Alison Heppenstall, Mingshu Wang, Urska Demsar, Rob Lemmens, and Jing Yao.
This contribution underwent peer review based on a full paper submission.

Lessons from spatial transcriptomics and computational geography in mapping the transcriptome

Alexis Comber [1, 2], Eleftherios Zormpas [3], Rachel Queen [3], and Simon J. Cockell [3]

[1] School of Geography, University of Leeds, Leeds, UK
[2] Leeds Institute for Data Analytics, University of Leeds, UK
[3] Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK

Correspondence: Alexis Comber (a.comber@leeds.ac.uk)

Abstract. Spatial data, data with some form of location attached, are the norm: all data are spatial now. However spatial data requires consideration of three critical characteristics, observation spatial auto-correlated, process spatially non-stationarity and the effect of the MAUP. Geographers are familiar with these and have tools, rubrics and workflows to accommodate them and understand their impacts on statical inference, understanding and prediction. However, increasingly researchers in non geographical domains, with no experience of, or exposure to quantitative geography or GIScience are undertaking analyses of such data without full or any understanding of the impacts of these spatial data properties. This short paper describes recent interactions and work with research in gene analysis and Spatial Transcriptomics, and highlight the opportunities for GIScience to inform and steer the many new users of spatial data.

Keywords. Spatial data, Molecular Biology, GIScience, Spatial autocorrelation, the MAUP, Process spatial non-stationarity

1 Introduction

All data are spatial - they are collected collected somewhere and come with some form of locational information attached to them. Location may be the latitude and longitude of an observation directly captured using a GPS-enabled device or it may be indirectly collected at a particular facility as some-one taps in to the transport system with a travel card.

The endemic characteristics of spatial data and analyses of spatial data are:

1. Observations are frequently auto-correlated across space, violating assumptions of independence and randomness in classic statistics
2. Many processes, when examined spatially, exhibit spatially non-stationarity (i.e. have local relationships over space). This requires explicit consideration of location in statistical models.
3. All spatial data are affected by their spatial support – the spatial scale over which it was collected or aggregated. This MAUP (Modifiable Areal Unit Problem) effect causes distortions in statical models, relationships and other inferences.

These characteristics and issues are present in ALL spatial data but unknown within many of the research communities, who are ubiquitously collecting and analysing spatial data (from agriculture to zoology, via climate science and sustainability). The danger of not explicitly accounting for spatial data characteristics, are erroneous inference (understanding and prediction) and decision making.

Many disciplines have been brought into the world of spatial data and analysis through the ease of spatial data creation and collection and the ease of analysis using powerful, free open source GIS software. These have frequently addressed problems at scales that are familiar to geographers: landscape, neighbourhood, small area, agricultural field parcel, administrative area, regular remote sensing and modelling grid, etc. A typical workflow is in these disciplines is that their analyses are correctly undertaken if they have some geographic input, but not always if they do not: Comber et al. (2015) documents many instances where researchers seeking to "optimise" the spatial configuration of land based renewable energy using a location-allocation approach, had misspecified the algorithm relative to the stated problem and pressed the wrong button the GIS. As a result there have been a number clarion calls from computational geographers for other research communities to pay heed to these issues, to avail themselves of the many nuanced open source tools and toolkits for undertaking spatial analysis, and to adopt transparent and reproducible practices in their analyses of spatial data

(Brunsdon and Comber, 2021; Comber and Wulder, 2019; Brunsdon and Comber, 2020; Brunsdon, 2016; Franklin, 2023; Comber and Harris, 2022; Nüst et al., 2018; Nüst and Eglen, 2021). There is some recent evidence the some of non-geographical research communities investigating processes at these scales, using spatial data are responding to this (e.g. Thorson et al. (2023); Zhao et al. (2024); Sa'adi et al. (2023)). However, some disciplines using spatial data operate in a very different way, without strong traditions of openness, sharing and reproducibility) and work with data at very different spatial scales that are unfamiliar to geographers.

One such domain is Bioinformatics and specifically genome analysis undertaken using a relatively recent technology: Spatial Transcriptomics.

2 Spatial Transcriptomics

The traits of a cell (its phenotype) are determined by its Transcriptome, the protein-coding of the organism's genome. The spatial patterns of gene expression provide indicators of the molecular biology of the cells that comprise the tissue and thus of tissue function. Recent developments in methods for *in-situ* hybridisation (ISH) have resulted in the ability to comprehensively examine gene expression spatially. This is Spatial Transcriptomics, (ST) in which intact tissue section is probed (i.e. stretches of single-stranded RNA are identified) at discrete locations within the cell ("spots") and the RNA is associated with the closest spot, supporting RNA analysis over specific locations. ST combines tissue imaging with comprehensive transcriptome quantification, with analyses undertaken at increasingly finer spot resolutions (from 100 μm to 500 nm), and now supporting sub-cellular analyses.

A typical ST problem is shown in Figure 1 which shows gene expression data from a 10X Genomics Visium experiment with the histological image, as included in the `Spaniel` R package (Queen et al., 2019).

The ability to undertake RNA sequencing and to examine the spatial patterns of gene expression with ST has resulted in it being highlighted as a 'method of the year' by Nature Methods in 2021 (Marx, 2021) and ST technology has since gained significant interest in the wider field of molecular biology. For example, the Human Cell Atlas is an international project that aims to profile every cell in the human body using ST technology to create large cellular maps (Rozenblatt-Rosen et al., 2017). However, the current state of the science in ST data analysis is naive from a spatial data science perspective: it does not take account of critical considerations in spatial data analysis, and does not take full analytical advantage of the opportunities afforded by location. Often a simple (a-spatial) clustering to group similar observations is the only analysis undertaken, with little use made of the spatial information except to map the clusters. This is evidenced in a number of recent key ST review, methods and advances papers (Marx, 2021; Noel et al., 2022; Dries et al., 2021; Rao et al., 2021): the world of ST is working with spatial data but is completely unaware of the 60+ years of tool development and method refinement in computational and quantitative geography. In one instance, ST researchers, observing spatial autocorrelation, suggested the need for more spatially aware analyses of the variation in gene expression (Svensson et al., 2018). They proposed an approach (SpatialDE) that measures spatial dependence by "testing whether gene expression levels at different locations covary in a manner that depends on their relative location", without any reference to existing methods such as Moran's I, Anselin's LISA, Geary's C, Getis and Ord's G and so on. This is typical of the ST domain, where researchers are not inherently inter- and cross-disciplinary, and method development does not look outside of the ST / Bioinformatics domains.

Some recent work has started to promote robust and open spatial data science approaches within ST. It has undertaken a number of initial investigations using tools from computational geography and is starting to link computational geographers and bioinformaticians working with ST. It has identified a key deficiencies in current ST approaches, a summary of which from Zormpas et al. (2023) is shown in Figure 2, and has identified opportunities for ST that have not yet been fully exploited within ST, some which are described below. More recent work has also suggested opportunities for computational geography to explore processes at the micro-scale and in 3D.

3 Current ST Opportunities

Currently, spatial information in mainstream ST analyses is used in only very limited ways:

1. To cluster gene expression at spot locations.
2. To examine specific cell and or gene positioning
3. To evaluate variations in gene expression, for between clusters, located over space.

Typical ST workflows treat at best location quite arbitrarily, at worst ignore it completely.

Example 1: the activation state of an immune cell is determined by examining its neighbourhood. Sub-cellular observations are aggregated in different ways (grid, image-segmentation etc) to at least cell level. The choice of aggregation unit will influence process understanding through scale (MAUP) driven distortions but this is completely unknown and ignored.

Example 2: the typical approach for clustering gene expression constructs a graph weighted by gene expression similarity (actually the principal components from a PCA) using a user-specified number of nearest neighbours. Graph-based partitioning methods are used to identify homogeneous sub-graph regions, about which biolog-

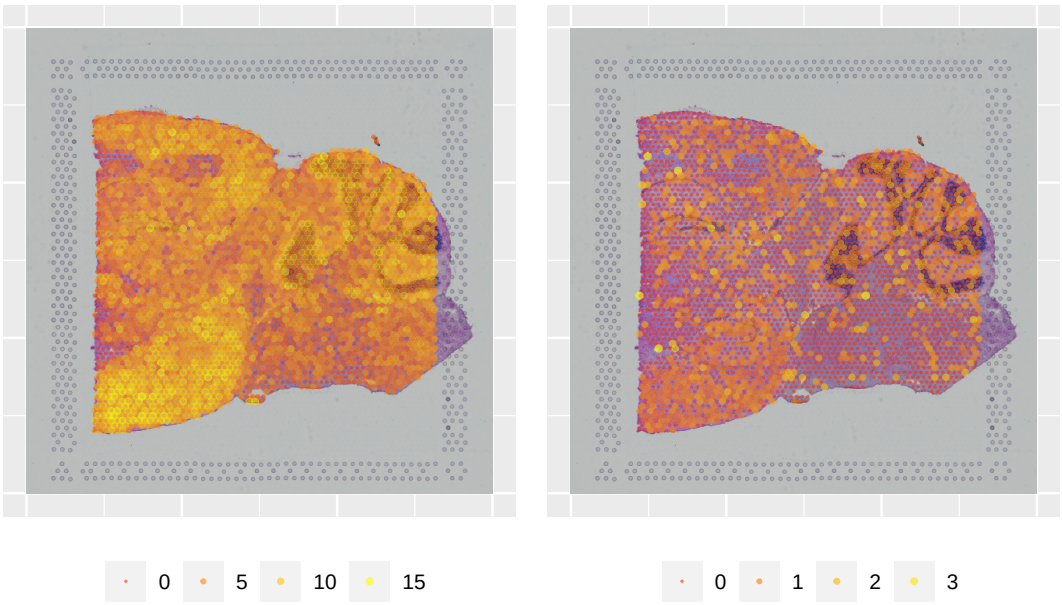

Figure 1. An example of Spatial Transcriptomics gene expression data, plotted using the Spaniel package: counts of all genes per spot (right) and of ENSMUSG00000024843.

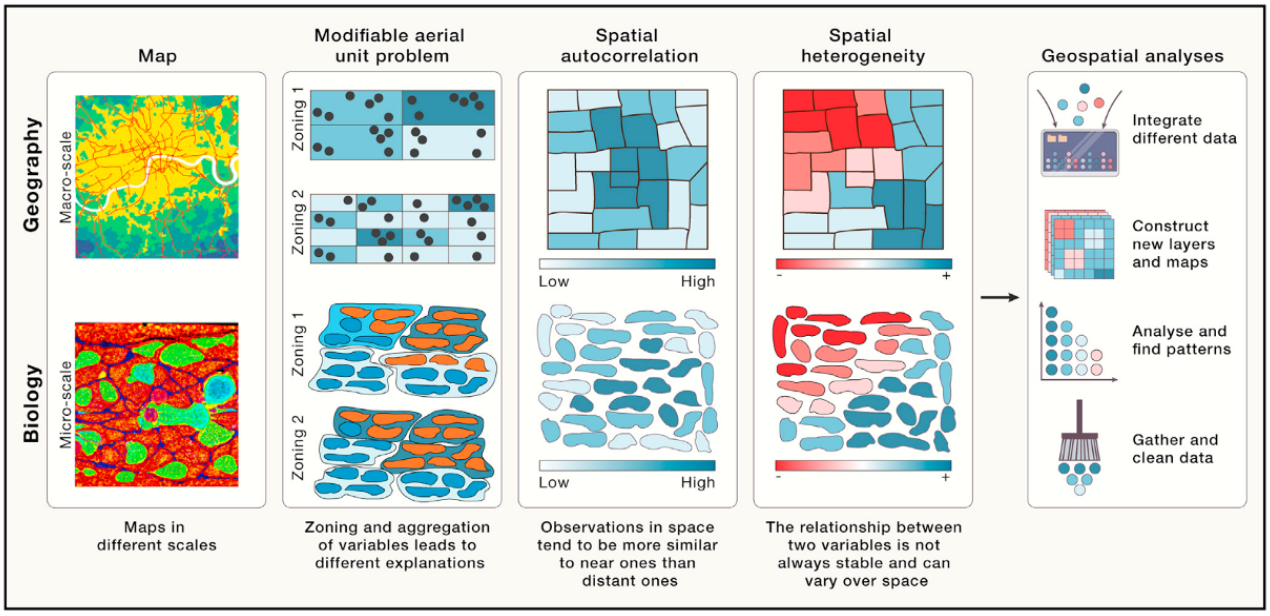

Figure 2. The parallel inferential and analysis opprotunities for Spatial Transcriptomics arising from computational geography methods and paradigms (figure from Zormpas et al (2023).

ical processes are inferred from their spatial pattern and distribution.

Effectively such workflows firstly treat observations as discrete and ignores their relative locations. Treating data independently in this way risks missing important information that can be extracted from location.

There are a number of opportunities to leverage thinking, concepts, tools and approaches from computational geography to enhance the ST analyses that are currently undertaken. These include:

- Spatial Gene Expression Patterns: Investigate and visualise spatial patterns of gene expression across tissues to identify spatially regulated genes and understand their functional implications.

- Cellular Heterogeneity: Explore the heterogeneity within tissues by identifying distinct cell populations and understanding their spatial distributions (this has a central role in cancer research and the investigation of the tumour micro-environment).

- Spatial Interaction Networks: how different cell types or genes interact spatially, providing insights into the local cellular environment.
- Functional Annotation: Connect spatial information with functional annotation, helping to decipher the biological relevance of gene expression patterns in specific locations.
- Pathway Analysis: Explore spatially enriched biological pathways, shedding light on the functional significance of gene expression patterns within specific regions.

In some of these activities, location is being include but only as a covariate and in others it is being used but not in an spatially informed or transparent way.

4 Future ST Opportunities

In a similar way there are also opportunities to leverage computational geography to develop *new* ST approaches, that to support wider and deeper analyses of the spatial pattern of gene expression within cell and tissue and extend standard techniques from from computational geography and spatial data handling into ST. These include:

- Spatial Autocorrelation metrics such as Local Indicators of Spatial Association (LISA) statistics could be used to identify SA, i.e. spatial clusters of genes with similar expression patterns. This would identify and detect local patterns of spatial autocorrelation and pinpoint regions with significant clustering. Different SA detection techniques identify different kinds of SA: from its simple presence in Moran's I, to local measures LISA statistics, and the identification of high and low value clustering in Getis and Ord's G-statistic. Some gene toolkits are using these approaches e.g. MERINGUE, but not in a way that takes advantage of the latest CG developments both in terms of data structures, and toolkits.
- Spatial Error models could be used to simulate SA by incorporating a spatial error term in order to test for the probability that unobserved factors are influencing gene expression in spatially correlated ways. Such factors potentially include protein levels, chromatin structure, molecules that activate or inhibit certain pathways, or even sets of genes whose expression is relative constant with the result at they are omitted from downstream analyses because they are not sufficiently interesting (i.e. they don't pass the high variance filter).
- Geostatistical approaches can be used to continuous construct surfaces of gene expression over unsampled locations, by interpolating observed expression patterns at sampled locations. There are many CG techniques for continuous surface construct from point observations: from kriging (gold standard but complex) to IDW (good enough and quick). Such approaches may be useful in situations where spots have large distances between them (e.g. Visium).
- Hotspot and Coldspot detection would also be supported by methods for geostatistical and SA analyses. These would identify localised regions with unusually high or low gene expression levels, helping to pinpoint biologically relevant areas for further investigation.
- Spatial Overlay can be used to integrate gene expression data with other spatial resolved information. In a manner similar to a GIS analysis, layers representing tissue structures or landmarks (i.e., central veins in the liver) could be used to examine the spatial relationships between gene expression patterns and specific anatomical features. Buffer Analyses can be incorporated in the spatial overlay operations above, allowing for zones around sampled locations to analyse the impact of nearby geographical features on gene expression, helping to elucidate the influence of the local micro-environment on spatial transcriptomics.

The above is not an exhaustive list but these are standard techniques in computational geography, often undertaken in a GUI GIS by practitioners, but more commonly using R or Python packages.

5 And what about Computational Geography?

There are also opportunities for quantitative and computational geography to learn from the domain of spatial transcriptomics. Much of computational geography operates at landscape related scales: from river catchments to census areas, and many of the rubrics we are familiar with including the various "laws" of spatial dependence / distance decay and spatial heterogeneity / non-stationarity (Tobler, 1970; Goodchild, 2004), are empirical in origin: they were developed and tested through empirical observation of the world we live in, which is not normally or even randomly distributed. The theory, reasoning and logical came later. Thus they are also grounded in landscape scale spatial processes and their measurement mostly captured in 2D, with some occasional extensions into the z-dimension (height, depth or elevation) but more frequently into time to consider spatio-temporal processes.

Exploring new domains at new scales provides opportunities for quantitative and computational geography to develop new understandings and knowledge of spatial processes and behaviours at previously unexplored micro-scales, and potentially techniques in micro-scale 3D. this inter-disciplinary turn allows current accepted geographical wisdoms and accepted paradigms to be examined and tested at finer scales. For example, how relationships between process grain, spatial sampling (support) and scale

distortions manifest themselves at sub-cellular levels. Or how to extend spatially informed statistical regression models into 3D. These have the potential to stimulate new and generalisable insights into interactions between (spatial) sampling frameworks and processes.

6 Conclusions

This is a not a standard research paper. But its message has relevance to the broader GIScience, Quantitative and Computational Geography community:

- All data are spatial and many researchers from other domains are increasingly working with spatial data.
- But as we know, spatial data and geographic process are subject to the MAUP, observation spatial autocorrelation and process non-stationarity.
- This makes spatial data analysis different from data analysis.
- In some cases new communities using spatial data are doing it well and have recognised the benefits of working with us, but in others they have not.
- Thus there are opportunities for computational geographers and GI scientists to enhance research activities in important domains and applications that are working with spatial data (especially the ones that society deems to be important and funds more highly like gene transcription analysis).

The contribution that the GIScience and computational geography community can make are to create environments for robust analyses of spatial data, ones that are informed by observation spatial autocorrelation, an expectation of process spatial non-stationarity (heterogeneity) in statistical model outcomes, and consideration of MAUP effects.

Such activities could, for example include the development of route maps for spatial regression models. This would take the naive user from basic whole map / whole transcriptome regression models, regression with spatial dummies (fixed effects), through spatial econometric models, multi-level models, GWR and MGWR in the manner of Comber et al. (2023) and into recent work with space and space-time GAMs (Comber et al., 2024). They could create guides to encourage users to really consider the nature of any spatially varying processes that are found with such tools. Spatially vary processes may be found due to a number of confounding reasons, including a poor conceptual model, the lack of universal laws do not govern gene behaviours, bad measurements with locational bias, or unaccounted for local factors. Or they could be due to a truly spatially varying process that exhibit process spatial heterogeneity. Finally, methods to explore MAUP effects could be illustrated. Despite the MAUP being a core consideration in all analyses of spatial data, its effects are rarely tested for (including in much geographic research). However, there are well established approaches for quantifying the impact of the MAUP and for determining appropriate sampling and aggregation scales as documented in Comber and Harris (2022). It is essential to make spatial data users aware of these issues and to provide rubrics, tools and workflows to support them and their work.

Acknowledgements. This research is in part supported by a studentship from the UK Medical Research Council (MRC) Discovery Medicine North (DiMeN) Doctoral Training Partnership (grant number: MR/ N013840/1), and the University of Leeds by supporting the lead author with a period of Study Leave.

References

Brunsdon, C.: Quantitative methods I: Reproducible research and quantitative geography, Progress in Human Geography, 40, 687–696, 2016.

Brunsdon, C. and Comber, A.: Big issues for big data: challenges for critical spatial data analytics, arXiv preprint arXiv:2007.11281, 2020.

Brunsdon, C. and Comber, A.: Opening practice: supporting reproducibility and critical spatial data science, Journal of Geographical Systems, 23, 477–496, 2021.

Comber, A. and Harris, P.: The importance of scale and the MAUP for robust ecosystem service evaluations and landscape decisions, Land, 11, 399, 2022.

Comber, A. and Wulder, M.: Considering spatiotemporal processes in big data analysis: Insights from remote sensing of land cover and land use, 2019.

Comber, A., Dickie, J., Jarvis, C., Phillips, M., and Tansey, K.: Locating bioenergy facilities using a modified GIS-based location–allocation-algorithm: Considering the spatial distribution of resource supply, Applied Energy, 154, 309–316, 2015.

Comber, A., Brunsdon, C., Charlton, M., Dong, G., Harris, R., Lu, B., Lü, Y., Murakami, D., Nakaya, T., Wang, Y., et al.: A route map for successful applications of geographically weighted regression, Geographical Analysis, 55, 155–178, 2023.

Comber, A., Harris, P., and Brunsdon, C.: Multiscale spatially varying coefficient modelling using a Geographical Gaussian Process GAM, International Journal of Geographical Information Science, 38, 27–47, 2024.

Dries, R., Chen, J., Del Rossi, N., Khan, M. M., Sistig, A., and Yuan, G.-C.: Advances in spatial transcriptomic data analysis, Genome research, 31, 1706–1718, 2021.

Franklin, R.: Quantitative methods II: Big theory, Progress in Human Geography, 47, 178–186, 2023.

Goodchild, M. F.: The validity and usefulness of laws in geographic information science and geography, Annals of the Association of American Geographers, 94, 300–303, 2004.

Marx, V.: Method of the Year: spatially resolved transcriptomics, Nature methods, 18, 9–14, 2021.

Noel, T., Wang, Q. S., Greka, A., and Marshall, J. L.: Principles of spatial transcriptomics analysis: a practical walk-through in kidney tissue, Frontiers in Physiology, 12, 2317, 2022.

Nüst, D. and Eglen, S. J.: CODECHECK: an Open Science initiative for the independent execution of computations underlying research articles during peer review to improve reproducibility, F1000Research, 10, 2021.

Nüst, D., Granell, C., Hofer, B., Konkol, M., Ostermann, F. O., Sileryte, R., and Cerutti, V.: Reproducible research and GIScience: an evaluation using AGILE conference papers, PeerJ, 6, e5072, 2018.

Queen, R., Cheung, K., Lisgo, S., Coxhead, J., and Cockell, S.: Spaniel: analysis and interactive sharing of spatial transcriptomics data, bioRxiv, p. 619197, 2019.

Rao, A., Barkley, D., França, G. S., and Yanai, I.: Exploring tissue architecture using spatial transcriptomics, Nature, 596, 211–220, 2021.

Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A., and Teichmann, S. A.: The Human Cell Atlas: from vision to reality, Nature, 550, 451–453, 2017.

Sa'adi, Z., Yaseen, Z. M., Farooque, A. A., Mohamad, N. A., Muhammad, M. K. I., and Iqbal, Z.: Long-term trend analysis of extreme climate in Sarawak tropical peatland under the influence of climate change, Weather and Climate Extremes, 40, 100 554, 2023.

Svensson, V., Teichmann, S. A., and Stegle, O.: SpatialDE: identification of spatially variable genes, Nature methods, 15, 343–346, 2018.

Thorson, J. T., Barnes, C. L., Friedman, S. T., Morano, J. L., and Siple, M. C.: Spatially varying coefficients can improve parsimony and descriptive power for species distribution models, Ecography, p. e06510, 2023.

Tobler, W. R.: A computer movie simulating urban growth in the Detroit region, Economic geography, 46, 234–240, 1970.

Zhao, X., Tan, S., Li, Y., Wu, H., and Wu, R.: Quantitative analysis of fractional vegetation cover in southern Sichuan urban agglomeration using optimal parameter geographic detector model, China, Ecological Indicators, 158, 111 529, 2024.

Zormpas, E., Queen, R., Comber, A., and Cockell, S. J.: Mapping the transcriptome: Realizing the full potential of spatial data analysis, Cell, 186, 5677–5689, 2023.