Teaching spatial data science

Ane Rahbek Vierø, IT University of Copenhagen & Michael Szell, IT University of Copenhagen, ISI Foundation, Complexity Science Hub Vienna, and Pioneer Centre for AI.

Spatial data science is an emerging field building on geographic information science, geography, and data science. Here we first discuss the definition and history of the field, arguing that it indeed warrants a new label. Then, we present the design of our course Geospatial Data Science at IT University of Copenhagen and discuss the importance of teaching not just spatial data science tools but also spatial and critical thinking. We conclude with a perspective on the potential future for spatial data science, arguing that qualitative theory and methods will continue to play an important role despite new GeoAI-related advances.

Keywords: spatial data science, data science, python, teaching, GeoAI

1. What is spatial data science?

Spatial data and spatial approaches are of great importance when addressing many of the core challenges of our time, such as the climate crisis, urban sustainability, and socio-spatial segregation (Leszczynski & Crampton, 2016; Tenkanen, 2024; UN, 2012). Moreover, a large chunk of the world's growing data collections requires some spatial understanding to process adequately (Batty, 2013; Huang & Wang, 2020; Leszczynski & Crampton, 2016). There is thus an increasing number of people and professions who work with and make use of spatial data – many however without a background in 'traditional' spatial disciplines such as geographic information science (GIS) or geography. The growing interest in and demand for spatial data analysis presents an urgent need to bridge disciplines, and to ensure that existing methods, tools, and know-how are understandable and findable for people outside of GIS. Meanwhile, new computing tools, growing data sets, and the increasing use of machine learning and AI means that traditional spatial domains have a lot to gain from closer integrating with other data science disciplines (S. Rey et al., 2023; Singleton & Arribas-Bel, 2021). These two trends might be part of the explanation for the increasing popularity of the term 'spatial data science'.

There is no universal definition of neither data science nor spatial data science – or of synonyms such as 'geospatial' or 'geographic' data science. Most however agree that data science exists "at the interface between computer science and statistics" (Arribas-Bel & Reades, 2018) and involves a combination of methods from math, statistics, and computer science, and importantly, also domain knowledge (Dodge, 2021) (Fig. 1).

Building on this definition, one could argue that spatial data science is about combining computational, statistical, and mathematical methods with spatial data to tackle spatial questions. Additionally, spatial data science also involves geovisualization (Çöltekin et al., 2018; Elwood, 2011) and, ideally, drawing on the insights and experience from the long history of geography (Arribas-Bel & Reades, 2018).

Some might argue that this definition is not very different from previous terms such as 'geocomputation', 'geoinformatics', etc. (Tenkanen, 2024). There can undoubtedly be some boosterism involved in the continuous invention of new labels, and discussions of e.g. spatial big data are far from new (Kitchin, 2014a). We nevertheless argue that spatial data science often is qualitatively different from earlier strands of quantitative and computational geographic analysis. Firstly, the past decades have witnessed a data deluge, with a rapid increase in the number and size of available data sets (Kitchin, 2014b), combined with the emergence of a range of new data sources, stemming from both the ubiquity of the internet, embedded sensors, and mobile devices (Arribas-Bel, 2014; Leszczynski & Crampton, 2016). This means that we not only have more data than before, but also completely different types of data, that, for example, allows us to analyze, model, and predict human movements at an unprecedented spatial scale and resolution. Secondly, the increasing use of machine learning and AI technologies, enabled by a large increase in computing powers, is fundamentally reshaping analytical approaches. This in turn affects which problems we address and what type of questions we ask. Finally, many of the people working with spatial data science are not from traditional geospatial backgrounds but have come to work with spatial data due to the strong spatial component in many new data sets, and in many of today's largest IT companies (Arribas-Bel & Reades, 2018; Kitchin, 2013). This development both creates exciting opportunities for interdisciplinarity and new perspectives, but also poses a challenge in ensuring that important experiences and insights from geography are carried forward.

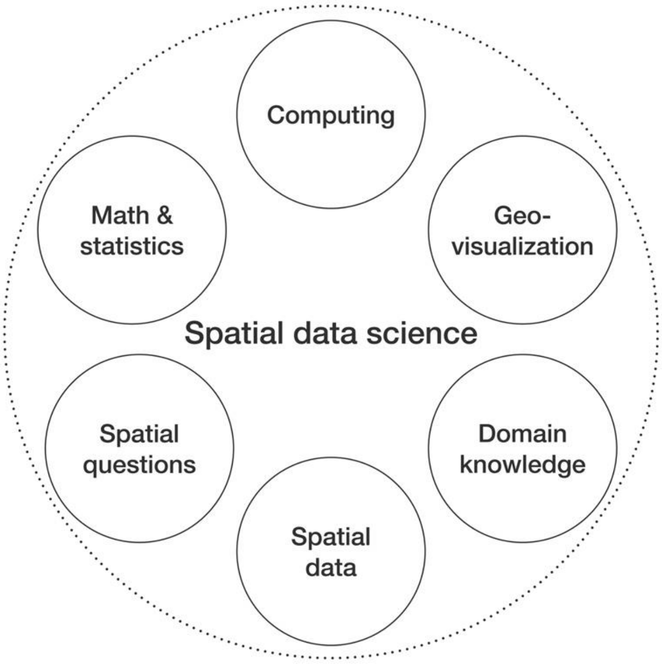

Figure 1: Spatial data science is an intersection of multiple disciplines. Inspired by (Dodge, 2021) in (Tenkanen, 2024).

2. Teaching spatial thinking to data science students

As a response to the growing need for people with knowledge of both data science and spatial methods, we have developed the course Geospatial Data Science at the IT University of Copenhagen, which we have been teaching since 2022. The course is offered as an elective as part of the master's program in Data Science but is also attended by students enrolled in Computer Science, as well as some students from other disciplines. The choice to develop a course dedicated specifically to the spatial elements of data science reflects similar developments at other universities, which have started either teaching data science to GIS and geography students, or teaching geographic theories and methods to data science students (MacLachlan & Dennett, 2022; Mann et al., 2023; S. Rey et al., 2023; Tenkanen, 2023; Tenkanen et al., 2023). While the course so far has been popular among students, who often are motivated by topics related to climate and sustainability, it is a challenge to establish the necessary foundations for working with spatial data in a 7.5 ECTS course over the span of 14 weeks.

In the course we cover both the mathematical and statistical methods behind core geospatial methods, introduce the most commonly used spatial Python libraries, and emphasize the importance of the visual elements of spatial data exploration, analysis, and communication. Course topics range from basic handling of spatial coordinates to analysis of spatial clustering, spatial autocorrelation, and spatial networks, with an emphasis on how the different methods can be applied to concrete examples (Fig. 2).

More importantly though, teaching spatial data science to students outside of traditional spatial disciplines necessitates teaching students how to think spatially. Learning to think spatially first of all means understanding the meaning behind Tobler's first law stating that "everything is related to everything else, but near things are more related than distant things" (Miller, 2004). Spatial thinking also means to appreciate why spatial variation must be considered in statistical analysis of spatial phenomena (S. Rey et al., 2020), and being familiar with potential pitfalls such as the modifiable areal unit problem, distortions from map projections, etc. More generally however, it is about understanding why place and space matters for many of the topics we study in data science. It is thus not just about correct handling of spatial coordinates, but about understanding how a spatial perspective might contribute with new insights and understandings when studying for example mobility, accessibility, economic inequity, or environmental degradation, to mention a few of the topics of the students' exam projects. In the course, we further emphasize the importance of reproducibility (i.e., all results should be documented and possible to reproduce with the accompanying code), as it is one of the important benefits offered by completely scripting-based spatial analysis.

The course thus aims to both 1) teach the technical expertise for Python-based exploration, analysis, and visualization of spatial data, 2) give the students the skills to apply critical, spatial perspectives on important topics related particularly to sustainability and urban planning. Although a one semester course can only teach a limited selection of all the many methods, perspectives, and tools with relevance to spatial data science, our experience is that after an intensive crash-course in spatial thinking, data science students can utilize their existing skills sets for impressive spatial data science applications.

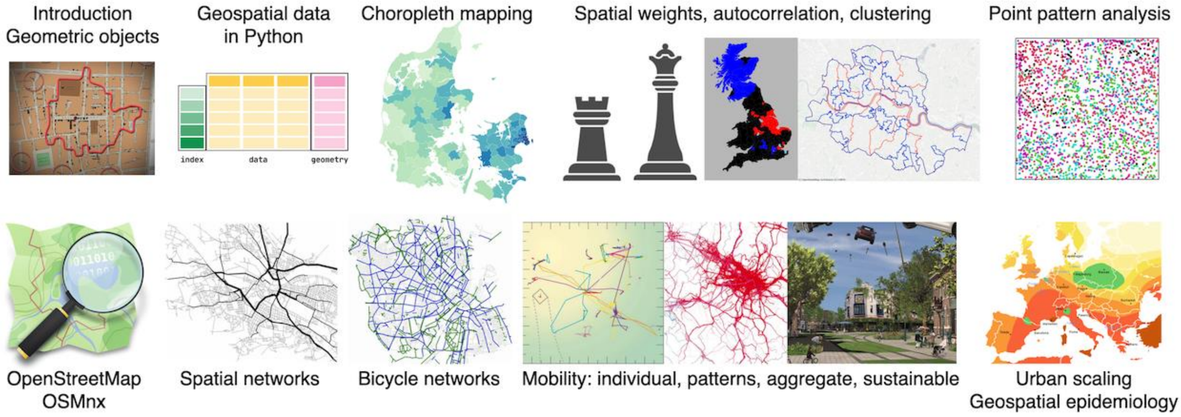

Figure 2: Overview of topics taught in our course Geospatial Data Science. Open teaching materials at: https://github.com/mszell/geospatialdatascience.

3. Next steps for spatial data science?

Although labels change and the term ‘spatial data science’ might be replaced in the future – especially given the rapid development of many data science tools and technologies – we believe there will be a continued need for people with strong data science skills combined with knowledge on how to apply them to spatial topics. For a future successful development of spatial data science, we need education that not only equips the students with technical skill set, but also incorporate domain knowledge, draw on the rich body of experiences and insights from geography and related disciplines, and make students aware of some of the ethical challenges that go along with new tools and data sources. The large growth in data from mobile devices has for example given discussions of (the lack of) spatial privacy a new urgency. Finally, spatial data science is per definition multidisciplinary, and it is thus important to break down silos and combine the strengths of all relevant fields.

Like many other disciplines, the proliferation of AI and ML will undoubtedly have a big influence on the development of spatial data science in the coming years. ML and AI, and the emerging sub-field of GeoAI (Janowicz et al., 2020) can potentially offer huge benefits to spatial data science, not least given the growing data sizes that can be hard to handle for other types of analytical tools. It is however important to remember that theory and insights from more qualitative methods surely will continue to play an important role, and that some of the expected outcomes from AI in many ways mimic earlier predictions of Big Data being the supposed “end of theory” and end of traditional science (Kitchin, 2013). It will thus continue to be important to teach core spatial analysis techniques to students in and outside of the spatial disciplines.

References

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, *49*, 45–53. https://doi.org/10.1016/j.apgeog.2013.09.012

Arribas-Bel, D., & Reades, J. (2018). Geography and computers: Past, present, and future. *Geography Compass*, *12*(10), e12403. https://doi.org/10.1111/gec3.12403

Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, *3*(3), 274–279. https://doi.org/10.1177/2043820613513390

Çöltekin, A., Janetzko, H., & Fabrikant, S. (2018). Geovisualization. *Geographic Information Science & Technology Body of Knowledge*, *2018*(Q2). https://doi.org/10.22224/gistbok/2018.2.6

Dodge, S. (2021). A Data Science Framework for Movement. *Geographical Analysis*, *53*(1), 92–112. https://doi.org/10.1111/gean.12212

Elwood, S. (2011). Geographic Information Science: Visualization, visual methods, and the geoweb. *Progress in Human Geography*, *35*(3), 401–408. https://doi.org/10.1177/0309132510374250

Huang, B., & Wang, J. (2020). Big spatial data for urban and environmental sustainability. *Geo-Spatial Information Science*, *23*(2), 125–140. https://doi.org/10.1080/10095020.2020.1754138

Janowicz, K., Gao, S., McKenzie, G., Hu, Y., & Bhaduri, B. (2020). GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, *34*(4), 625–636. https://doi.org/10.1080/13658816.2019.1684500

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262–267. https://doi.org/10.1177/2043820613513388

Kitchin, R. (2014a). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Kitchin, R. (2014b). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. SAGE Publications Ltd. https://doi.org/10.4135/9781473909472

Leszczynski, A., & Crampton, J. (2016). Introduction: Spatial Big Data and everyday life. *Big Data & Society*, *3*(2), 2053951716661366. https://doi.org/10.1177/2053951716661366

MacLachlan, A., & Dennett, A. (2022). An Applied Geographic Information Systems and Science Course in R. *Journal of Open Source Education*, *5*(50), 141. https://doi.org/10.21105/jose.00141

Mann, M., Chao, S., Graesser, J., & Feldman, N. (2023). *PyGIS - Open Source Spatial Programming & Remote Sensing—Python Open Source Spatial Programming & Remote Sensing*. https://pygis.io/docs/a_intro.html

Miller, H. J. (2004). Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers*, *94*(2), 284–289. https://doi.org/10.1111/j.1467-8306.2004.09402005.x

Rey, S., Arribas-Bel, D., & Wolf, L. J. (2023). *Geographic Data Science with Python* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429292507

Rey, S. J., Arribas-Bel, D., & Wolf, L. J. (2020). *Spatial Regression—Geographic Data Science with Python*. Geographic Data Science with Python. https://geographicdata.science/book/notebooks/11_regression.html

Singleton, A., & Arribas-Bel, D. (2021). Geographic Data Science. *Geographical Analysis*, *53*(1), 61–75. https://doi.org/10.1111/gean.12194

Tenkanen, H. (2023). *Spatial Analytics*. https://spatial-analytics.readthedocs.io/en/latest/index.html

Tenkanen, H. (2024). *What is Spatial Data Science?* Spatial Data Science for Sustainable Development. https://sustainability-gis.readthedocs.io/en/latest/index.html

Tenkanen, H., Heikinheimo, V., Aagesen, H. W., Fink, C., & Hasanzadeh, K. (2023). *Automating GIS Processes 2023*. https://autogis-site.readthedocs.io/en/latest/index.html

UN. (2012). *The Future We Want*. https://sustainabledevelopment.un.org/futurewewant.html