The Impact of Data Science on Geography: A Review with Optimization Algorithms

Roberto de Oliveira Machado[a]
Universidade Nova de 061, Lisboa, Portugal
roberto.machado@campus.fcsh.unl.pt

Keywords: Data Science, Geography, Optimization Algorithms, Supervised Learning, Systematic Review.

Abstract: We conducted a systematic review using the PRISMA methodology, analyzing 2,996 studies and synthesizing 41 to explore the evolution of data science and its integration into geography. Optimization algorithms were employed to enhance the efficiency and precision of literature selection. Our findings reveal that data science has evolved over five decades, facing challenges such as the integration of diverse spatial data and the increasing demand for advanced computational skills. In the field of geography, data science emphasizes interdisciplinary collaboration and methodological innovation. Techniques like large-scale spatial data analysis and predictive algorithms hold promise for applications in natural disaster management and transportation optimization, enabling faster and more effective responses. These advancements highlight data science's pivotal role in solving complex spatial problems. This study contributes to the application of optimization algorithms in systematic reviews and underscores the necessity for deeper integration of data science into geography. Key contributions include identifying challenges in managing heterogeneous spatial data and promoting advanced analytical capabilities. The intersection of data science and geography leads to significant improvements in disaster management and transportation efficiency, fostering more sustainable and impactful environmental solutions.

1 INTRODUCTION

Data science is an interdisciplinary field that integrates statistical analysis, machine learning, and computational techniques to extract significant insights from large and complex datasets. With the advent of big data, characterized by high volumes, velocity, and variety, data science has become essential in analyzing and interpreting vast amounts of information. Technologies such as cloud computing, artificial intelligence, and advanced algorithms facilitate data processing and analysis, enabling real-time insights and decision-making across various domains.

The application of data science in geography has immense potential to revolutionize the analysis of spatial data and address complex issues. Predictive algorithms and spatial analysis techniques can enhance natural disaster prediction, optimize transportation networks, and manage resources sustainably. These innovations provide precise and actionable information, improving the effectiveness of spatial interventions and fostering community resilience (Yue, 2016; Singleton, 2019). However, the integration of data science into geography faces challenges, such as the need to harmonize heterogeneous spatial data, the demand for interdisciplinary expertise, and the lack of specialized tools for advanced spatial analysis, which hinder its broader adoption in the field of geography.

This study examines the evolution of data science and its integration into geography through a systematic review. It employs the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology to ensure a structured and unbiased synthesis of relevant studies. Additionally, supervised learning techniques, such as logistic regression and Naïve Bayes algorithms, are employed to optimize the literature selection process, reducing manual workload and improving precision.

A bibliometric analysis is conducted using Bibliometrix to explore trends, knowledge gaps, and contributions across disciplines.

[a] https://orcid.org/0000-0002-8346-4155

Machado, R. O.
The Impact of Data Science on Geography: A Review with Optimization Algorithms.
DOI: 10.5220/0013464200003935

In *Proceedings of the 11th International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM 2025)*, pages 219-230
ISBN: 978-989-758-741-2; ISSN: 2184-500X

The primary objectives of this work are to: (i) conduct a systematic review following the PRISMA protocol, enhanced by optimization algorithms, to ensure a comprehensive analysis of existing knowledge; (ii) examine the evolution of data science and its applications in geography; and (iii) provide actionable insights into how data science can be leveraged to address complex spatial challenges.

By achieving these objectives, the research highlights the transformative potential of data science in addressing spatial challenges and advancing interdisciplinary collaboration.

The article is structured into five sections: (i) systematic literature review; (ii) bibliometric analysis; (iii) results; (iv) discussion; and (v) conclusion.

2 METHODS

To ensure a structured and unbiased analysis of the literature, the study adhered to the PRISMA methodology (Pickering et al., 2014; Teh et al., 2020), which comprises four key stages: identification, screening, eligibility, and inclusion. Relevant studies were retrieved through comprehensive database searches in Scopus and Web of Science (WOS), employing carefully selected search terms to capture the intersection of data science and geography. Titles and abstracts were screened to exclude irrelevant articles based on predefined criteria, ensuring both relevance and methodological rigor. Full-text articles were subsequently assessed for eligibility, and only those meeting all criteria were included in the final analysis.

The screening process was enhanced through the implementation of supervised learning algorithms, such as Logistic Regression and Naïve Bayes. These algorithms automated key stages of the process, significantly reducing manual effort. Initially, additional classifiers, including Support Vector Machines, K-Nearest Neighbors, and Random Forest, were also evaluated. However, they were excluded due to their lower performance in terms of accuracy and false-negative rates. Logistic Regression exhibited the lowest false-negative rate, ensuring the inclusion of high-quality and relevant studies, while Naïve Bayes excelled in computational efficiency and generalization capability, making them suitable for handling the available dataset.

A bibliometric analysis conducted using Bibliometrix provided deeper insights into trends, knowledge gaps, and interdisciplinary contributions. Graphs and tables were generated in Excel, while spatial data visualizations were created in Jupyter Notebook, ensuring a robust and versatile approach to data analysis.

To enhance transparency and reproducibility, rigorous inclusion and exclusion criteria were applied throughout the process. Potential biases were mitigated through manual verification (Ólafsdóttir & Tverijonaite, 2018; Saltz & Dewar, 2019; Saltz & Krasteva, 2022) of algorithmic outputs and sensitivity analyses (Kantardzic, 2019) to assess the robustness of the findings.

2.1 Systematic Literature Review

The review followed six steps: formulation, identification, selection, eligibility, inclusion, and categorization. Two research questions were established: “What is the state of the art in data science?” and “What is its applicability in geography?” To ensure a structured selection process, specific inclusion criteria were defined, considering articles, conference papers, reviews, books, and book chapters that contained the relevant search terms in the title, abstract, and keywords, were aligned with the research questions, and were written in English. Conversely, exclusion criteria were applied to filter out editorials, notes, short surveys, data papers, and any studies misaligned with the research questions or written in languages other than English.

The literature search was conducted in the Scopus and Web of Science (WOS) databases, using the keywords “Data Science” AND “Geography” OR “Spatial” OR “Big Data” OR “Geographic” OR “Geospatial”. The search encompassed all available studies up to May 2023, focusing on abstracts to ensure relevance. The results were exported in CSV format for Scopus and Research Information Systems (RIS) format for WOS. A normalization process was then applied to merge and structure data from different sources into a unified CSV format, leveraging Python 3.8 and the Pandas library for efficient data processing.

The next phase involved the manual classification of a set of articles into relevant and irrelevant categories. This step was followed by the implementation of a supervised active learning cycle, leveraging the following libraries: Pandas; sklearn.feature_extraction.text.TfidfVectorizer; sklearn.linear_model.LogisticRegression; sklearn.naive_bayes.MultinomialNB; sklearn.metrics.recall_score; sklearn.metrics.confusion_matrix.

The Term Frequency - Inverse Document Frequency (TF-IDF) technique was used to measure the importance of a word in a document, considering a collection of documents or corpus. This measurement was performed by assigning a weight during the text vectorization process (Bafna et al., 2016; Chen et al., 2016; Sah et al., 2020). The TF-IDF calculation for the term 't' in a document 'd' is defined as shown in Equation (1):

$$TF - IDF(t, d, D) = TF(t, d) * IDF\ (t, D) \quad (1)$$

where:

- $TF(t, d)$ corresponds to the number of times the term t appears in the document d;
- $IDF\ (t, D)$ represents the inverse document frequency of the term t in the set of documents D, calculated as: $IDF(t, D) = log\ (N/(1 + n_t)$

where:

- N is the total number of documents in the set;
- n_t is the number of documents that contain the term t.

Here is the total number of documents in the set and is the number of documents that include the term t. This method assigns higher weights to less frequent terms and lower weights to more frequent terms (O'Mara-Eves et al., 2015; Chen et al., 2016).

In the selection procedure, the Logistic Regression classifier demonstrated superior performance in classification, identifying higher proportion of relevant studies compared to the total number of relevant articles detected. This indicates that Logistic Regression had a lower false-negative rate, being more effective in identifying relevant studies. The performance metrics of active learning for the two algorithms used (Logistic Regression and Naïve Bayes) in categorizing the relevance of studies related to data science concepts. The recall value was calculated using Equation (2):

$$\frac{TP}{TP + FN} \quad (2)$$

where:

- TP represents true positives (relevant studies correctly identified) and FN represents false negatives (relevant studies incorrectly marked as irrelevant) (O'Mara-Eves et al., 2015).

Throughout the cycle, the binary Logistic Regression classifier was used to predict the category of a dependent variable based on the values of the independent variables. The result of this process was 0 for "irrelevant" or 1 for "relevant", as shown in Equation (3):

$$P(x_i, \theta) = \frac{1}{1 + e^{-z}} \quad (3)$$

where:

- $P(x_i, \theta)$ is the probability that observation x_i belongs to class y_i;
- e is the constant Euler's number, approximately 2.71828;
- z is the linear combination of features weighted by parameters: $z = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip}$;
- $\frac{1}{1+e^{-z}}$ is the sigmoid function, which transforms the linear combination z into a value between 0 and 1, interpreted as the probability of belonging to the positive class.

Implementing a systematic literature review using supervised learning can reduce the time required to identify relevant studies. However, this method is not without biases, potentially omitting up to 5% of relevant literature (Ros et al., 2017; Yu & Menzies, 2019; Ferdinands et al., 2020; Wang et al., 2020; van De Schoot et al., 2021). To mitigate this, the process followed guidelines proposed by Yang (2018), which recommend finalizing the review process after identifying 50 irrelevant records.

After completing the screening process with the Logistic Regression classifier, an additional selection was conducted using the Naïve Bayes classifier. The Naïve Bayes classifier employs a probabilistic classification mechanism based on Bayes' Theorem (Yang, 2018), expressed by the following Equation (4) (Zhang & Gao, 2011; Jurafsky & Martin, 2019):

$$P(y|x_1, x_2, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \quad (4)$$

where:

- $P(y|x_1, x_2, \ldots, x_n)$ is the conditional probability of class y given the variables $x_1, x_2, \ldots, x_n$;
- $P(y)$corresponds to the prior probability of class y;
- $P(x_i|y)$ represents the conditional probability of observing the feature x_i given that the class is y;
- $\prod_{i=1}^{n} P$ indicates the multiplication of all conditional probabilities for each feature x_i;
- $\propto$ = 3.822 is a hyperparameter used to improve classifier performance.

The results were compared to identify potential

omissions of relevant literature. Following this, a full reading of the selected studies was conducted, with particular attention given to frequently cited authors not included in earlier phases. Relevant studies were integrated to minimize potential biases. In the eligibility phase, studies misaligned with the research questions were excluded from the review process.

In the final categorization phase, information from eligible studies was extracted. Bibliometric data were analyzed using the open-source tool Bibliometrix, which provided disaggregation by space, time, thematic area, and journals. Graphs and tables were generated in Excel, and spatial data visualizations were developed using Jupyter Notebook. It is important to acknowledge that the review was conducted by a single reviewer, which could introduce subjectivity in the inclusion and exclusion criteria. Additionally, only textual studies were considered due to limitations in applying text classifier algorithms to digitized documents. The Python code for the systematic review is well-documented and ensured for reproducibility. It is available for consultation on [GitHub] and [Google Colab].

3 RESULTS

The results of this systematic review provide a comprehensive overview of the main trends and challenges in applying data science to geography. The initial database search identified 2,996 studies. After removing duplicates, 1,806 studies were selected for screening. Title and abstract analysis led to the selection of 126 documents for full reading. After examination, 35 studies were included, while the rest were excluded for not aligning with the research question. Citation analysis contributed six additional studies, including three books and three other documents, two of which were in image format. The total number of studies for categorization and synthesis was 41.

3.1 Bibliometric Analysis

This systematic review includes studies from 10 countries. The United States leads with 24 studies, followed by the Netherlands with four, the United Kingdom and China with three each, and Australia with two. The remaining countries contributed one publication each. The reviewed studies span from 1966 to 2022, with the earliest published in 1966 and the latest in December 2022. The most representative years were 2017 (17%), followed by 2018 (15%), and 2016 and 2022 (10% each). Until 2012, studies focused on the contribution of statistics and computing to the foundations of data science. Between 2013 and 2017, research addressed the field's definition and applicability. From 2018 to 2022, the literature expanded to various topics, from specific methodologies to guidelines for knowledge extraction, with a link to geography.

Among the six most representative areas of knowledge, computer science is the most prominent, accounting for 35% of occurrences, followed by social sciences and decision sciences, each at 16%. Mathematics accounts for 14%, earth and planetary sciences for 11%, and business, management, and accounting for 8%. Regarding the distribution of study types, 46% are articles, 25% are reviews, 17% are conference papers, and 12% are books.

The most frequent periodicals include *Communications of the ACM*, *Geographical Analysis*, *Geography Compass*, and *Proceedings of the National Academy of Sciences*, each with two publications. Among the top authors, van der Aalst W. leads with three studies, while Smyth P., Arribas-Bel D., and Cao L. each have two. The six most cited studies cover topics such as data mining (Fayyad et al., 1996), process mining (van der Aalst, 2016), the intersection of supply chain management and data science (Waller & Fawcett, 2013),, the challenge of defining data science and its link to big data (Provost & Fawcett, 2013), the role of data scientists (Davenport & Patil, 2012), and the distinction between data science and statistics (Dhar, 2013).

3.2 Definition of the Concept of Data Science

As a starting point, we provide an overview of the concept of data science as defined in previous studies. Data science is broadly conceived as a flexible and practical approach to extracting knowledge or useful patterns from data (Dhar, 2013; Provost & Fawcett, 2013; Kelleher & Tierney, 2018; Vicario, G., & Coleman, 2019). It combines three fundamental components: computing, statistics (Vicario, G., & Coleman, 2019; Blei, 2017; Cao, 2018; Yan & Davis, 2019), and domain knowledge (Song & Zhu, 2015; Muller et al., 2019). Computing facilitates large-scale parallel computing, algorithms, and data mining to uncover useful information (Provost & Fawcett, 2013; Kelleher, J., & Tierney, 2018; Donoho, 2017). Statistics provide a quantitative framework, applying concepts such as mean, standard deviation, hypothesis testing, and statistical inference (Waller, & Fawcett, 2013). Domain knowledge is crucial for

analyzing and extracting knowledge pertinent to specific areas (Muller et al., 2019).

Vicario et al. (2019), Yan & Davis (2019), and van der Aalst (2016) emphasize the interdisciplinary nature of data science, where various disciplines collaborate to extract knowledge from data. In contrast, Song & Zhu (2019) describe data science as multidisciplinary, with each field contributing uniquely while operating within a common context. Cao (2018) further classifies data science as transdisciplinary, where different domains converge to extract relevant information from data. Therefore, data science can be inter-, multi-, or transdisciplinary depending on the project's scope, the individuals involved, and the evolving nature of the field. Interdisciplinary and multidisciplinary approaches keep data science closely linked to its foundational areas, while achieving a transdisciplinary state requires robust research that encompasses both interdisciplinary and multidisciplinary elements.

The Venn diagram illustrates the convergence of disciplines in data science, emphasizing its interdisciplinary nature, as noted by Blei & Smyth (2017). The overlap of computer science and domain knowledge produces "advanced analysis," where models are created, and analyses are conducted. The intersection of mathematics, statistics, and computer science enables "machine learning" through data and algorithms. The overlap of domain knowledge with mathematics and statistics leads to "traditional analysis," using classical methods to identify patterns (Provost & Fawcett, 2013; Kelleher & Tierney, 2018). At the core, where all three disciplines intersect, lies the discovery of knowledge through integrated data analysis, statistical methods, and advanced computing.

Song & Zhu propose a distinct perspective, suggesting that data science comprises three pillars: data, technologies, and people. They argue that the third pillar —skilled professionals — is essential for selecting appropriate data and applying advanced technologies in analysis. Blei & Smyth (2017) also emphasize the human element in data science, highlighting the need for data scientists to understand the problem domain, devise inferential methods, and effectively use computational tools. Scheider et al. (2020) further note that such skills are often acquired through professional interactions and practical implementation, rather than solely through academic training (Hicks & Irizarry, 2018).

The proliferation of data, driven by numerous human activities and technological advancements, represents a significant revolution in knowledge production (van der Aalst, 2016). Data science has emerged as a flexible field capable of analyzing large volumes of structured and unstructured data.

Structured data, typically stored in SQL, are tabulated for easier storage and processing, while unstructured data, in NoSQL systems, are more irregular and complex to analyze (Kelleher & Tierney, 2018). To manage this complexity, data scientists often create an *n*m* analytical matrix where *'n'* represents entities (rows) and *'m'* represents attributes (columns). This matrix, incorporating data from multiple sources such as databases, data warehouses (Kelleher & Tierney, 2018), big data lakes (Carniel & Schneider, 2021), online content, and social media streams, is fundamental to data science projects. Significant time and resources are spent organizing, cleaning, and updating this matrix.

3.3 Origin and Evolution of the Concept

In the 1960s, statistician John Tukey and computer scientist Peter Naur laid the foundations of data science, leading many to view it as a descendant of statistics and computer science (Blei & Smyth, 2017). While incorporating methods and approaches from these fields, data science has evolved to meet contemporary analytical needs driven by the omnipresence of data collected ubiquitously (van der Aalst & Damiani, 2015). Over the past two decades, this evolution has consolidated data science as a distinct science.

The term “data science” first appeared in literature in the preface of Peter Naur's book *Concise Survey of Computer Methods* in the 1970s, where he defined it as the "science of dealing with data" (Cao, 2018, p. 9). However, John Tukey's earlier work in 1962 proposed incorporating data analysis processes within statistics, advocating for the inclusion of new theories, computers, visualization tools, and larger datasets (van der Aalst, 2016; Cao, 2018). This need for change was echoed in the field of computing.

In 1966, Peter Naur introduced “Datalogy” (Naur, 1966), the systematic study of data, including representation and automatic processing mechanisms. Naur characterized this field as broad and interdisciplinary, essential for tasks across different domains and data types. Despite its significance, the role of statistics in defining data science foundations remains debated, as statisticians traditionally focused more on formulating theories than addressing practical challenges associated with large datasets and computational complexities (van der Aalst, 2016).

In the 1970s, John Tukey's *Exploratory Data*

Analysis (Tukey, 1997) introduced techniques aimed at improving results through hypothesis testing, unbiased data, and conventional statistics, which later came to be known as confirmatory data analysis. Concurrently, the relational data model SQL revolutionized database information storage, indexing, and retrieval (Kelleher & Tierney, 2018), laying the groundwork for "Knowledge Discovery in Databases" (KDD) in the late 1980s (Cao, 2018). The seminal 1996 article "Data Mining and Knowledge Discovery in Databases" by Fayyad et al. (1996) outlined a five-stage process for knowledge discovery: selection, pre-processing, transformation, data mining, and pattern interpretation.

Building on the KDD model, other methodologies emerged both in academia and industry due to the growing demand for data analysis tools. Notable among these was the Cross-Industry Standard Process for Data Mining (CRISP-DM), organized around a six-stage lifecycle, providing a flexible and iterative framework for data mining projects (Martínez-Plumed, 2021). As data science gained prominence in the 1990s, it became increasingly associated with both research and practical applications (Zhu & Xiong, 2015).

The early 21st century witnessed further advancements, including Cleveland's action plan for teaching data science and the launch of NoSQL databases, propelling the field forward. The increasing of systems led to the rise of big data, creating the need for new processing frameworks such as MapReduce and Hadoop (van der Aalst, 2016; Cao, 2018). MapReduce, with its Map and Reduce functions, and Hadoop, integrating Hadoop Distributed File System (HDFS) with MapReduce, enabled efficient storage and distribution of large data volumes across clusters (Kumar & Mohbey, 2022).

In the second decade of the 21st century, data science became more clearly defined, integrating computer science, mathematics, statistics, and domain-specific knowledge (Brady, 2019). The role of the data scientist, as highlighted by Davenport & Patil (2012), gained prominence as data science became critical for problem-solving and deriving insights. This period saw an increase in publications on data science infrastructure and processes, such as van der Aalst's *Process Mining Data Science in Action* (2016) and Donoho's concept of "greater data science" activities (2017). The profession of data scientist emerged as both essential and desirable, although discussions on the competencies required for these professionals continue to evolve in literature and academia (Vicario & Coleman, 2019).

3.4 Function and Applicability

Our analysis demonstrates that data science has solidified its position as a key discipline for solving complex problems through the integration of statistical and computational methods (Cleveland, 2001; Waller & Fawcett, 2013; Martínez-Plumed et al., 2021). In addition to addressing the challenges posed by the big data phenomenon (Song, & Zhu, 2015), data science aims to generate both actional and actionable knowledge (Donoho, 2017). It employs automated analyses to understand various phenomena, converting data into actionable insights, diagnostics, and automated decisions (Provost & Fawcett, 2013; van der Aalst, 2017). This transformation follows a methodology that ranges from data collection to knowledge acquisition (Cao, 2018). As Brady (2019) notes, data science employs specific processes to explore and describe data, identify meaningful patterns, and present them effectively, adapting and redefining the traditional objectives of statistics and computer science. Van der Aalst & Damiani (2015) emphasize that data science is capable of addressing four fundamental questions: "What happened?" "Why did it happen?" "What will happen?" and "What is the best decision to make?" Notably, although data science is often associated with large datasets, it is equally effective when applied to smaller datasets, which can also yield valuable information for solving specific problems.

Most studies suggest that the knowledge discovery process in data science follows a lifecycle encompassing systematic methodologies, open and interdisciplinary approaches, and synergies from various disciplines (Song, & Zhu, 2015; van der Aalst & Damiani; 2015; Cao, 2017; Yu & Kumbier, 2020). However, there is some discrepancy concerning the number of stages in this cycle, ranging from four (van der Aalst, 2017) to eight stages (Yu & Kumbier, 2020). The most referenced model is CRISP-DM (Provost & Fawcett, 2013; Song & Zhu, 2015; Martínez-Plumed, 2021). Its main strength lies in its independence from specific software or analysis techniques (Kelleher, & Tierney, 2018). CRISP-DM, an extension of the KDD model (Martínez-Plumed, 2021), is structured into six stages: (i) understanding the project objectives and defining the problem; (ii) collecting data for analysis; (iii) performing cleaning actions to correct or remove inaccurate or irrelevant information; (iv) applying algorithms to identify useful patterns in the data and establish representative models; (v) evaluating the model and the developed stages; and (vi) implementing the model, assessing its practical viability and ability to generate knowledge

or value from the data (Larose & Larose, 2014). It is important to note that the CRISP-DM process is iterative, not necessarily following a linear sequence of stages.

The technologies used in data science vary across organizations and projects (Kelleher & Tierney, 2018). As the size of the organization or the volume of data increases, so does the complexity of the ecosystem to be implemented. A typical data science ecosystem comprises tools and components from multiple software vendors, capable of processing data in diverse formats. Commonly used tools include SQL and NoSQL databases (van der Aalst & Damiani, 2015), MapReduce for distributed computing (Blei & Smyth, 2017), and Hadoop for storage, along with associated components such as HBase, Hive, Pig, Mahout, Storm, and Spark (Song, & Zhu, 2015; van der Aalst & Damiani, 2015; Kumar & Mohbey, 2022; Abdalla, 2022). These components are critical during the data representation and computation phase (Donoho, 2017). The programming languages SQL, Python, and R are employed for querying, visualization, data cleaning, and calculations (Brunner & Kim, 2016).

Among the various tools available, Jupyter Notebook has gained significant prominence as an open-source web-based interactive development environment. Jupyter Notebook stands out for its interface, which facilitates communication, abstraction, and data manipulation (Carniel & Schneider, 2021). It enables users to access databases, configure, and organize workflows for scientific computing and machine learning with ease. Moreover, Jupyter Notebook is highly flexible, allowing seamless integration of extensions, such as Python and R libraries, for analysis, model application, and result visualization. This flexibility and ease of use contribute to a reduction of the time required for analysis. One of the tool's most valuable features is its ability to integrate both text and code in a single location, a concept known as "literate programming". This is particularly beneficial in data science projects involving multidisciplinary teams, where not all members may possess a deep understanding of all tasks, thereby making integrated documentation an essential feature.

3.5 Relationship with Geography

Data science emerges in geography as a complementary, though somewhat limited, process (Singleton & Arribas-Bel, 2019). A review of the literature analysis reveals no consensus on the definition of the term, with various designations, including "earth data science," "geospatial data science," "spatial data science," and "geographic data science."

Yue et al. (2016, p. 84) define "earth data science" as "the ability and knowledge to apply the data science paradigm to EO data". Xie et al. (2017, p. 3) describe "geospatial data science" as "the science of data-driven approaches focusing on the geospatial domain". Carniel et al. (2021, p. 1) define "spatial data science" as a "subclass of object-oriented data science and spatial data". Finally, Andrienko et al. (2017, p. 15) refer to "geographic data science" as a field for working "with data that incorporates spatial and often temporal elements".

Spatial data science is defined as the discipline that explores data-driven approaches, with a particular emphasis on the spatial domain. Therefore, this review adopts the term "spatial data science", which, according to Singleton Singleton & Arribas-Bel (2019) and Bowlick & Wright (2018), has the potential to revitalize quantitative geography and geocomputation methodologies. This revitalization occurs through the application of advanced techniques for collecting, preparing, exploring, and visualizing large datasets, alongside the use of machine learning and spatial statistical methods to detect patterns and model spatial dependencies (Xie et al., 2017). Spatial data science seeks to extract valuable information about space (Andrienko et al., 2017), time, and distance (Carniel, & Schneider, 2021). High-performance computational procedures, algorithms, and specific programming languages are essential for extracting this information within an interdisciplinary or transdisciplinary framework.

One of the most compelling applications of data science in geography is the integration of EO data with socio-economic datasets. Recent studies have demonstrated how combining EO data with social media and economic data enhances our understanding of socio-economic-environmental systems, thus facilitating improved urban planning and disaster management (Yue et al., 2016).

Visualization tools are essential in geographic data science. For instance, the Urban Space Explorer tool uses a coordinated multiple-view interface to analyze social media posts in relation to population distribution and points of interest. This tool has proven invaluable to urban planners, providing insights into urban dynamics and supporting data-driven decision-making (Xie et al., 2017).

Xie et al. (2017) view spatial data science as transdisciplinary, addressing the limitations of isolated data mining and machine learning approaches by integrating four key elements:

mathematics, statistics, computer science, and geospatial techniques. Palomino et al. (2017) argue that spatial data science is inherently interdisciplinary, emerging from the intersection of three distinct fields: Geographic Information Science, data science, and CyberGIS, which is defined as "GIS based on advanced cyberinfrastructure and e-science" (VoPham et al., 2018, p. 2). Carniel & Schneider (2021) also regard spatial data science as interdisciplinary, rooted in three core disciplines: mathematics/statistics, computer science, and business information/domain specialization. However, the literature lacks clarity on the specific techniques employed within the subclasses that emerge from the intersection of these core disciplines in spatial data science.

The integration of the three core disciplines of data science — computer science, mathematics and statistics, and domain knowledge —, can be effectively visualized using a Venn diagram. In the context of mathematics and statistics, techniques such as Kriging, bootstrap, and autocorrelation are employed. In computer science, programming and managing large spatial datasets are crucial, utilizing tools such as Spatial Hadoop and specialized open-source libraries like PySal, PyMove, Geopandas, and Moving Pandas. In domain knowledge, comprehending the spatial context — including location, distance, and spatial and temporal interactions — is crucial.

The implementation of data science projects within spatial contexts is often underemphasized in the literature. Palomino et al. (2017) and VoPham et al. (2018) advocate for workflows utilizing open-source and cloud computing technologies. Palomino et al. (2017) propose a comprehensive framework consisting of two complementary models: a four-stage model—setting up the work environment, data management (collection, cleaning, and transformation), analysis and visualization, and data publication, with an emphasis on collaborative work—and an iterative six-stage workflow, which builds on the previous model by adding specific steps for data import, processing, cleaning, transformation, visualization, modeling, and communication of results. Carniel & Schneider (2021) introduce a life cycle based on conventional data science workflows, comprising four components: problem definition and understanding, spatial data modeling and understanding, spatial reasoning, and spatial data visualization.

In practice, spatial data science benefits from the inherent flexibility in implementing data science projects. While establishing models, such as CRISP-DM, are widely used, scientists retain the flexibility to adapt the lifecycle or workflow to suit the specific project context (Martínez-Plumed et al., 2021). The skills required for spatial data science continue to evolve, with teamwork, machine learning techniques, cloud computing, and proficiency in programming languages such as Python, R, and SQL being crucial (Jiang & Chen, 2021).

Successfully executing a spatial data science project requires proficiency in three key areas: domain knowledge, mathematical and statistical skills, and computer science (Carniel & Schneider, 2021; Bowlick & Wright, 2018). However, the role of quantification in geography remains a subject of ongoing debate, in part due to the fact that many practitioners lacking a strong background in statistics and computing. Increasing emphasis on these areas is crucial for strengthening the connection between geography and data science, both academically and professionally.

There is a longstanding tradition of using open-source programming languages such as Python and R, in spatial analyses (Anselin & Rey, 2022). The CyberGIS domain advocates for the application of cloud computing in geography (Palomino et al., 2017). Specialized software, such as QGIS and ARCGIS, has adapted to this trend by integrating open-source extensions for querying, analyzing, and visualizing data. In recent decades, geocomputation and remote sensing have increasingly incorporated large datasets, algorithms, and open-source code into their analyses (Arribas-Bel & Reades, 2018). Geographers are also actively involved in smart city projects, which presents new opportunities for analyzing large datasets (Scheider et al., 2020).

In these environments, structured and unstructured data from diverse sources are frequently employed for complex analyses (Brady, 2019). This convergence of factors, coupled with ongoing developments, can cultivate a stronger relationship between data science and geography. To date, this relationship has remained fragmented and indirect (Singleton, & Arribas-Bel, 2019). Should this relationship deepen, it could mark a new chapter for the discipline, reinforcing its position in a data-driven world, open science (Palomino et al., 2017), and cloud computing. This scenario is consistent with contemporary data science approaches to data collection, transformation, processing, and analysis.

4 DISCUSSION

4.1 Systematic Review

This study employed an efficient, transparent, and reproducible process for bibliographic selection, providing a detailed and technical overview of the methodologies and algorithms used. Mathematical equations and detailed discussions on algorithm performance were provided, along with well-documented Python code that can be easily adapted and implemented both locally and on web-based platforms.

However, several limitations were encountered in implementing Python pipelines for continuous bibliography updates. The Scopus and Web of Science (WOS) APIs imposed functional restrictions, limiting the number of requests and records that could be retrieved and increasing query complexity. Consequently, article selection required direct interaction with the platforms. Furthermore, WOS does not allow direct CSV data export, requiring the conversion of RIS files to CSV using Python. This process introduces potential errors and inconsistencies, requiring rigorous standardization and cleaning procedures to ensure data quality.

4.2 Data Science

Data science plays a pivotal role in addressing complex problems by utilizing statistical and computational methodologies. In the era of big data, characterized by increasing volume, velocity, and variety, data science has evolved to automate and enhance data analysis processes, thereby improving diagnostics, predictive analytics, and decision-making tools.

By redefining traditional statistical and computational objectives, data science generates actionable insights across various domains, highlighting its transformative potential in addressing contemporary challenges.

4.3 Application in Geography

In geography, data science emerges as a complementary yet underutilized tool. The lack of consensus surrounding terms such as "earth data science," "geospatial data science," and "spatial data science" highlights the need for standardized terminology. Among these, "spatial data science" has gained prominence as a promising framework to rejuvenate quantitative geography and geocomputation methodologies.

Advanced techniques, such as machine learning and predictive analytics, offer significant potential across spatial domains. For instance, predictive algorithms enhance natural disaster management by providing more accurate forecasts and facilitated effective response strategies. The analysis of large datasets optimizes transportation routes, reduces congestion, and improves logistical efficiency. Furthermore, data science plays a crucial role in sustainable resource management by identifying patterns and trends that foster the adoption of environmentally friendly practices.

4.4 Challenges and Ethical Considerations

The application of data science in geography presents significant challenges, including the integration of heterogeneous spatial data, the need for advanced computational and statistical expertise, and ethical concerns surrounding data privacy and usage. However, spatial data science offers innovative solutions by integrating mathematics, statistics, computer science, and geospatial techniques. Tools such as open-source software and cloud computing provide promising pathways for addressing these challenges.

4.5 Interdisciplinary Collaboration

Interdisciplinary collaboration among geographers, data scientists, statisticians, and ethicists is crucial for addressing complex challenges at the intersection of geography and data science. Previous studies have shown that multidisciplinary teams optimize the analysis of large spatial datasets. Recommended strategies include forming diverse teams, fostering advanced data analysis skills, and adhering to rigorous ethical standards to promote methodological innovation and ensure sustainable solutions.

4.6 Limitations and Future Directions

Future research should prioritize the integration of advanced data science techniques into geography to tackle complex spatial challenges. Key areas include leveraging machine learning algorithms and incorporating novel data sources, such as Internet of Things (IoT) sensors, for real-time dynamic analyses. However, the adoption of these techniques is hindered by the limited proficiency of many geographers in programming and computational tools, which restricts the effective application of these methods. Bridging this gap will necessitate

interdisciplinary collaboration and targeted training programs that align technological advancements with practical needs.

Emerging technologies, including artificial intelligence and cloud computing, hold transformative potential for spatial research. These technologies enhance data processing capabilities, facilitate the integration of large datasets, and improve analytical precision. To fully capitalize on these benefits, it is essential to address skill gaps and develop user-friendly tools tailored to non-specialists. By overcoming these challenges, we can enable innovative, sustainable solutions for critical geographic issues, such as disaster management, urban planning, and resource optimization.

5 CONCLUSIONS

This study highlights the application of supervised learning algorithms, specifically Logistic Regression and Naïve Bayes, to optimize the systematic review process. These techniques notably increased efficiency and precision, while reducing the time required for literature selection, thereby demonstrating the transformative potential of data science methodologies in systematic reviews. By showcasing these improvements, the study illustrates how data science approaches can effectively enhance various research processes.

The investigation reveals that the majority of data science studies were conducted in the United States of America. This is largely attributed to country's advanced technological infrastructure, prestigious academic institutions, substantial financial investment in research, supportive policies for technological development, and the presence of major technology companies that drive innovation and the application of data science. In the field of geography, this infrastructure and support have facilitated the precise and efficient analysis of large spatial datasets. Advanced methodologies, including machine learning and data mining, hold significant potential to improve natural disaster management, optimize transportation routes, and promote sustainable resource management.

However, several significant challenges remain. The need for advanced computational and statistical skills, alongside ongoing ethical concerns regarding privacy and data usage, continues to present obstacles. Overcoming these challenges requires an interdisciplinary approach, integrating expertise from various domains to develop robust and ethical solutions.

This study highlights several practical implications for applied geography:

1. Natural Disaster Management: Predictive algorithms can substantially improve the accuracy of natural disaster forecasts, enabling more effective and timely evacuation plans and resource allocation. Enhanced disaster management can save lives, reduce economic losses, and strengthen community resilience.
2. Transportation Optimization: Data science techniques can facilitate the design of more efficient transportation networks, reducing congestion and improving logistics. Optimized transportation can lead to better urban mobility, reduced environmental impact, and a higher quality of life for city residents.
3. Sustainable Resource Management: By identifying patterns and trends in resource use, data science can promote sustainable practices. Effective resource management ensures the better conservation of natural resources and supports environmental sustainability.

In summary, data science offers new perspectives and solutions to complex geographical challenges. This study underscores the importance of interdisciplinary collaboration and methodological innovation in harnessing the transformative potential of data science within geography, offering a solid foundation for future research and practical applications.

ACKNOWLEDGMENT

This work was supported by the Fundação para a Ciência e Tecnologia [UI/BD/151395/2021].

REFERENCES

Abdalla, H. B. (2022). A brief survey on big data: technologies, terminologies and data-intensive applications. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-022-00659-3

Andrienko, G., Andrienko, N., & Weibel, R. (2017). Geographic Data Science. IEEE Computer Graphics and Applications, 37(5), 15–17. https://doi.org/10.1109/mcg.2017.3621219

Anselin, L., & Rey, S. J. (2022). Open-source software for spatial data science. Geographical Analysis, 54(3), 429–438. https://doi.org/10.1111/gean.12339

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

Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 61-66). IEEE. http://dx.doi.org/10.1109/ICEEOT.2016.7754750

Blei, D. M., & Smyth, P. (2017). Science and data science. Proceedings of the National Academy of Sciences of the United States of America, 114(33), 8689–8692. https://doi.org/10.1073/pnas.1702076114

Bowlick, F. J., & Wright, D. J. (2018). Digital Data-Centric Geography: Implications for Geography's Frontier. The Professional Geographer, 70(4), 687–694. https://doi.org/10.1080/00330124.2018.1443478

Brady, H. E. (2019). The challenge of big data and data science. Annual Review of Political Science, 22(1), 297–323. https://doi.org/10.1146/annurev-polisci-090216-023229

Brunner, R., & Kim, E. (2016). Teaching data science. Procedia Computer Science, 80, 1947–1956. https://doi.org/10.1016/j.procs.2016.05.513

Cao, L. (2017). Data science: A comprehensive overview. ACM Computing Surveys, 50(3), 1–42. https://doi.org/10.1145/3076253

Cao, L. (2018). Data Science Thinking: the next scientific, technological and economic revolution. https://www.amazon.com/Data-Science-Thinking-Scientific-Technological/dp/3319950916

Carniel, A., & Schneider, M. (2021). A Survey of Fuzzy Approaches in Spatial Data Science. In 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1-6. IEEE. https://doi.org/10.1109/FUZZ45933.2021.9494437

Chen, K., Zhang, Z., Long, J., & Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Systems with Applications, 66, 245-260. https://doi.org/10.1016/j.eswa.2016.09.009

Cleveland, W. S. (2001). Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. International Statistical Review, 69(1), 21–26. https://doi.org/10.1111/j.1751-5823.2001.tb00477.x

Davenport, T. & Patil, T. (2012). Data scientist: The sexiest job of the 21st century. Harvard business review, 90(10), 70-76. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64–73. https://doi.org/10.1145/2500499

Donoho, D. L. (2017). 50 years of data science. Journal of Computational and Graphical Statistics, 26(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. Ai Magazine, 17(3), 37–54. https://doi.org/10.1609/aimag.v17i3.1230

Ferdinands, G., Schram, R., de Bruin, J., Bagheri, A., Oberski, D. L., Tummers, L., & van de Schoot, R. (2020). Active learning for screening prioritization in systematic reviews - A simulation study. https://doi.org/10.31219/osf.io/w6qbg

Hicks, S. C., & Irizarry, R. A. (2018). A guide to teaching data science. The American Statistician, 72(4), 382–391. https://doi.org/10.1080/00031305.2017.1356747

Jiang, H., & Chen, C. (2021). Data Science Skills and Graduate Certificates: A Quantitative Text analysis. Journal of Computer Information Systems, 62(3), 463–479. https://doi.org/10.1080/08874417.2020.1852628

Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing. Pearson.

Kantardzic, M. (2019). Data mining. https://doi.org/10.1002/9781119516057

Kelleher, J., & Tierney, B. (2018). Data science. MIT Press.

Kumar, S., & Mohbey, K. K. (2022). A review on big data based parallel and distributed approaches of pattern mining. Journal of King Saud University - Computer and Information Sciences, 34(5), 1639–1662. https://doi.org/10.1016/j.jksuci.2019.09.006

Larose, D., & Larose, C. (2014). Discovering knowledge in data: an introduction to data mining (Vol. 4). John Wiley & Sons.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. A. (2021). CRISP-DM Twenty years Later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061. https://doi.org/10.1109/tkde.2019.2962680

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., ... & Erickson, T. (2019). How data science workers work with data: Discovery, capture, curation, design, creation. In Proceedings of the 2019 CHI conference on human factors in computing systems, 1-15. https://doi.org/10.1145/3290605.3300356

Naur, P. (1966). The science of datalogy. Communications of the ACM, 9(7), 485. https://doi.org/10.1145/365719.366510

Ólafsdóttir, R., & Tverijonaite, E. (2018). Geotourism: A Systematic Literature Review. Geosciences, 8(7), 234. https://doi.org/10.3390/geosciences8070234

O'Mara-Eves, A., Thomas, J., McNaught, J. et al. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic Reviews, 4, 5. https://doi.org/10.1186/2046-4053-4-5

Palomino, J., Muellerklein, O., & Kelly, M. (2017). A review of the emergent ecosystem of collaborative geospatial tools for addressing environmental challenges. Computers, Environment and Urban Systems, 65, 79–92. https://doi.org/10.1016/j.compenvurbsys.2017.05.003

Pickering, C., Grignon, J., Steven, R., Guitart, D., & Byrne, J. (2014). Publishing not perishing: how research students transition from novice to knowledgeable using systematic quantitative literature reviews. Studies in Higher Education, 40(10), 1756–1769. https://doi.org/10.1080/03075079.2014.914907

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data, 1(1), 51–59. https://doi.org/10.1089/big.2013.1508

Ros, R., Bjarnason, E., Runeson P. (2017). A Machine Learning Approach for Semi-Automated Search and Selection in Literature Studies. In Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering (EASE'17). Association for Computing Machinery, New York, NY, USA, 118–127. https://doi.org/10.1145/3084226.3084243

Saltz, J. S., Dewar, N. (2019). Data science ethical considerations: a systematic literature review and proposed project framework. Ethics and Information Technology, 21, 197–208. https://doi.org/10.1007/s10676-019-09502-5

Saltz, J. S., & Krasteva, I. (2022). Current approaches for executing big data science projects—a systematic literature review. PeerJ. Computer Science, 8, e862. https://doi.org/10.7717/peerj-cs.862

Scheider, S., Nyamsuren, E., Kruiger, H., & Xu, H. (2020). Why geographic data science is not a science. Geography Compass, 14(11). https://doi.org/10.1111/gec3.12537

Shah, K., Patel, H., Sanghvi, D. J., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5(1). https://doi.org/10.1007/s41133-020-00032-0

Singleton, A., & Arribas-Bel, D. (2019). Geographic Data Science. Geographical Analysis, 53(1), 61–75. https://doi.org/10.1111/gean.12194

Song, I., & Zhu, Y. (2015). Big data and data science: what should we teach? Expert Systems, 33(4), 364–373. https://doi.org/10.1111/exsy.12130

Teh, H. Y., Kempa-Liehr, A. W., & Wang, K. I. (2020). Sensor data quality: a systematic review. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-0285-1

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley Publishing Company.

van de Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. Nature Machine Intelligence, 3(2), 125–133. https://doi.org/10.1038/s42256-020-00287-7

van der Aalst, W. (2016). Process Mining: Data Science in action. Springer.

van der Aalst, W. M. P. (2017). Responsible Data Science: Using event data in a “People friendly” manner. In Lecture notes in business information processing (pp. 3–28). https://doi.org/10.1007/978-3-319-62386-3_1

van der Aalst, W. M. P., & Damiani, E. (2015). Processes Meet Big Data: Connecting Data Science with Process Science. IEEE Transactions on Services Computing, 8(6), 810–819. https://doi.org/10.1109/tsc.2015.2493732

Vicario, G., & Coleman, S. (2019). A review of data science in business and industry and a future view. Applied Stochastic Models in Business and Industry, 36(1), 6–18. https://doi.org/10.1002/asmb.2488

VoPham, T., Hart, J. E., Laden, F., & Chiang, Y. (2018). Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. Environmental Health, 17(1). https://doi.org/10.1186/s12940-018-0386-x

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics, 34(2), 77–84. https://doi.org/10.1111/jbl.12010

Wang, Z., Nayfeh, T., Tetzlaff, J., O’Blenis, P., & Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. PLoS ONE, 15(1), e0227742. https://doi.org/10.1371/journal.pone.0227742

Xie, Y., Eftelioglu, E., Ali, R. Y., Tang, X., Yan, L., Doshi, R., & Shekhar, S. (2017). Transdisciplinary Foundations of Geospatial Data Science. ISPRS International Journal of Geo-information, 6(12), 395. https://doi.org/10.3390/ijgi6120395

Yan, D., & Davis, G. K. (2019). A first course in data science. Journal of Statistics Education, 27(2), 99–109. https://doi.org/10.1080/10691898.2019.1623136

Yang, F. J. (2018). An implementation of Naive Bayes classifier. In International Conference on Computational Science and Computational Intelligence (pp. 301-306). https://doi.org/10.1109/CSCI46756.2018.00065

Yu, B. & Kumbier, K. (2020). Veridical Data Science. In Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20). Association for Computing Machinery, New York, NY, USA, 4–5. https://doi.org/10.1073/pnas.190132611

Yu, Z., & Menzies, T. (2019). FAST2: An intelligent assistant for finding relevant papers. Expert Systems With Applications, 120, 57–71. https://doi.org/10.1016/j.eswa.2018.11.021

Yue, P., Ramachandran, R., Baumann, P., Khalsa, S. J. S., Deng, M., & Jiang, L. (2016). Recent activities in Earth data science [technical committees]. IEEE Geoscience and Remote Sensing Magazine, 4(4), 84-89. https://doi.org/10.1109/MGRS.2016.2600528

Zhang, W., & Gao, F. (2011). An improvement to naive bayes for text classification. Procedia Engineering, 15, 2160-2164. https://doi.org/10.1016/j.proeng.2011.08.404

Zhu, Y., & Xiong, Y. (2015). Towards data science. Data Science Journal, 14(0), 8. https://doi.org/10.5334/dsj-2015-008