Journal of Environmental Management 345 (2023) 118758

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

A health data led approach for assessing potential health benefits of green and blue spaces: Lessons from an Irish case study

Oludunsin Arodudu [a,b,*], Ronan Foley [b], Firouzeh Taghikhah [c], Michael Brennan [d], Gerald Mills [e], Tine Ningal [e]

[a] Department of Sustainable Resources Management, State University of New York, College of Environmental Science and Forestry, Syracuse, NY, USA
[b] Department of Geography, Rhetoric House, National University of Ireland Maynooth, Co. Kildare, Ireland
[c] Dicipline of Business Analytics, The University of Sydney, Sydney, Australia
[d] Eastern and Midland Regional Assembly, 3rd Floor North, Ballymun Civic Centre, Main Street, Ballymun, Dublin 9, Ireland
[e] School of Geography, Newman Building, Belfield, University College Dublin, Ireland

ARTICLE INFO

Handling Editor: Jason Michael Evans

Keywords:
Green space
Blue space
Human health
Data clustering techniques
Health data led approach
Environmental justice

ABSTRACT

Research producing evidence-based information on the health benefits of green and blue spaces often has within its design, the potential for inherent or implicit bias which can unconsciously orient the outcomes of such studies towards preconceived hypothesis. Many studies are situated in proximity to specific or generic green and blue spaces (hence, constituting *a green or blue space led approach*), others are conducted due to availability of green and blue space data (hence, *applying a green or blue space data led approach*), while other studies are shaped by particular interests in the association of particular health conditions with presence of, or engagements with green or blue spaces (hence, adopting *a health or health status led approach*). In order to tackle this bias and develop a more objective research design for studying associations between human health outcomes and green and blue spaces, this paper discussed the features of a methodological framework suitable for that purpose after an initial, year-long, exploratory Irish study. The innovative approach explored by this study (i.e., *the health-data led approach*) first identifies sample sites with good and poor health outcomes from available health data (using data clustering techniques) before examining the potential role of the presence of, or engagement with green and blue spaces in creating such health outcomes. By doing so, we argue that some of the bias associated with the other three listed methods can be reduced and even eliminated. Finally, we infer that the principles and paradigm adopted by the health data led approach can be applicable and effective in analyzing other sustainability problems beyond associations between human health outcomes and green and blue spaces (e.g., health, energy, food, income, environment and climate inequality and justice etc.). The possibility of this is also discussed within this paper.

1. Introduction

While there is considerable evidence-based research assuming or confirming the health benefits of nature particularly green and blue spaces (Frumkin, 2003; Völker and Kistemann, 2011; Calogiuri and Chroni, 2014; Gascon et al., 2015; Kindermann et al., 2021), a closer inspection of such studies reveal that they follow three broad approaches or paradigm. The first group of studies are situated in proximity to particular or generic green (e.g., riparian forest, grassland etc.) and/or blue spaces (e.g., beach, coasts, river, stream etc.) and can be described as those employing the *green and blue space-led* approach, e.g., Mitchell et al. (2011), Wheeler et al. (2012); Völker and Kistemann (2014). The primary source of bias for the green and blue space-led approach is mostly specific individual or group interest in the potential health impacts of particular or generic individual green and blue spaces (Völker and Kistemann, 2011; Bell et al., 2018). This bias may chiefly be a result of some emotional connection or historical attachment to green and blue space and/or their health impacts or

* Corresponding author. Department of Sustainable Resources Management, State University of New York, College of Environmental Science and Forestry, Syracuse, NY, USA.
E-mail addresses: oludunsinarodudu@gmail.com, otarodud@esf.edu (O. Arodudu), Ronan.Foley@mu.ie (R. Foley), firouzeh.taghikhah@sydney.edu.au (F. Taghikhah), mbrennan@emra.ie (M. Brennan), tine.ningal@ucd.ie (T. Ningal).

https://doi.org/10.1016/j.jenvman.2023.118758
Received 7 February 2023; Received in revised form 7 August 2023; Accepted 9 August 2023
Available online 8 September 2023
0301-4797/

short-term/long-term academic research interest (Völker and Kistemann, 2014; Foley and Kistemann, 2015). It could also sometimes (to a lesser extent) be a result of convenience, i.e., ease of availability of green and blue space data (White et al., 2010; Foley, 2015). A second strand of studies was primarily driven by the availability of green and blue space data, with an additional emphasis on green and blue space indicators for testing human health and wellbeing, e.g., Normalized Difference Vegetation Index (NDVI), quality of green or blue spaces, tree canopy or cover, tree-to-house ratio, the proportion of green and blue space to land or population, access and/or closeness to green and blue spaces etc. Such studies can be described as applying the *green and blue space data led* approach. Examples of previous studies employing this approach include Fuller et al., 2007, Wheeler et al. (2015), Amoly et al. (2014), Wilker et al. (2014); Smith et al. (2017). The major source of bias for the green and blue space data-led approach is convenience (i.e., ease of availability of green and blue space data) and specific expertise in handling and analyzing green and blue space data (Wheeler et al., 2015; Smith et al., 2017). A third approach includes studies designed to assess the impacts of particular or generic green and/or blue spaces on specific health conditions of interest. This can be described as the *health led or the health status led approach*. The main source of bias for health led or health status led approaches (like the green and blue space-led approach) is the specific individual or group interest in potential health impacts or health improvement benefits of particular or generic green and/or blue spaces (Fuller et al., 2007; Wilker et al., 2014). However, unlike the green and blue space-led approach, bias towards the health led, or health status led approach is more likely to be a result of short or long-term academic research interest rather than emotional connection or historical attachment (Amoly et al., 2014; Gascon et al., 2015). Bias towards the health led or health status led approach may also arise to a somewhat lesser extent due to the availability of health data or history/incidences of disease outbreaks in particular communities (Hartig et al., 2003; McFarlane et al., 2013).

Noteworthy however, is the fact that it is not impossible to have studies with features of more than one of the approaches described, e.g., Fuller et al. (2007); Amoly et al. (2014) both have features of both green/blue space data led, and health/health status led approaches. However, irrespective of the approach or approaches adopted by individual green and blue space and health studies, they still provide considerable depth of evidence that green and blue spaces have health and wellbeing benefits. Despite the depth of evidence that such studies provide, they however also contain, to a greater or lesser degree, some level of implicit or inherent bias, which influences the choice of the population sample and/or health cases chosen for assessment, and unconsciously orients such studies in the direction of conceived hypothesis, i.e., that green and/or blue spaces have potential or verifiable health benefits (Groenewegen et al., 2006; Foley et al., 2018a). Bearing this in mind, there is a need to consider how to avoid the bias associated with these three approaches. Avoiding such bias will help ascertain the validity of the potential health benefits of green and blue spaces in a wider range of circumstances (Lachowycz and Jones, 2013; Arodudu et al., 2017).

To do this, our study beamed the searchlight on health data to provide an alternative approach. This required developing a methodological framework that first identifies spatially referenced areas with good and poor human health outcomes (from collected/available health data sources) before testing if the presence of, or probable engagement with green and blue spaces play any role in the creation of such human health outcomes or not (Maas et al., 2006; Mitchell et al., 2015). We described this approach as the *health-data led* approach. This approach confirms and ascertains the potential or verifiable health benefits of green and blue spaces by relying on deductions from health data (Arodudu et al., 2017; Foley et al., 2018a). The approach does not rule out or invalidate the possible impacts of other important factors, i.e., confounding and effect-modifying variables affecting human health outcomes, e.g., gender, age, race/ethnicity, climate/climate change, policy, income level, socio-economic deprivation, etc. (Pearson et al., 2014; Smith et al., 2017). It, however, focuses on identifying the best sample sites to investigate the impact of green and blue spaces on human health using available health data. In order to operationalize the health data-led approach, we need to devise a framework (i.e., a combination of tools, procedures and methods/methodologies) that can be applied to achieve that. In response to this, this methodological framework paper primarily provides synthesized insights and suggestions on the composition or features of such health-data led approach applicable for avoiding previous inherent or implicit bias associated with assessing relationships between human health outcomes and green and blue space availability and/or contact, based on lessons from an initial, year-long exploration study. The secondary objective of the paper is to provide proof of concept of the new approach, hence the presentation of preliminary results and discussions of the one-year explorative case study in Section 3.

The methodological framework adopted for the initial one-year exploration case study identified places with good and poor human health outcomes by testing the capabilities of the Global Moran's I Spatial Statistics/Clustering algorithm and the Anselin Local Moran's I Spatial Statistics/Clustering algorithm (both from the ArcGIS software) on spatially referenced Irish health data at the national level. Health data applied as indicators of good and poor health outcomes (on which the two algorithms were tested) included spatially referenced Irish self-reported health, disability and mortality data (at the national level), while the green and blue space indicators tested by this study for the assessment of green and blue space and human health relationships included the Green Proportion Index (GPI), the Blue Proportion Index (BPI) and the Green and Blue Proportion Index (GBPI). See Section 2.2.3 for the description of the green and blue space indicators and how they were computed. The indicators of good and poor health outcomes were the dependent variables (spatially referenced Irish self-reported health, disability, and mortality data), while green and blue space indicators were the independent variables (GPI, BPI and GBPI). The relationship between good and poor health outcomes and green and blue spaces was confirmed via the use of scatter plots and the application of simple linear regression modelling techniques. The confounding/effect-modifying variable investigated in this study is socio-economic deprivation. This was done using two indices derived from the Irish context for measuring socio-economic deprivation, namely SAHRU (Small Area Health Research Unit) and Pobal HP (Haase and Pratschke) deprivation index. Investigating confounding/effect-modifying variables helps ascertain the impact of other factors other than presence and/or absence and/or contact with green and blue spaces on human health outcomes in every study or context. Alternative methodological options that can be applied for each step in the methodological framework are suggested in relevant sub-sections in Section 2. Excerpts and a full version of the technical report from the one-year exploration case study can be found in Section 3 (Foley et al., 2018a).

2. Methodology framework for a health-data led approach for investigating associations between green and blue spaces and human health outcomes

Due to the centrality of the identification of spatially referenced areas with good and/or poor human health outcomes to the health data led approach, of great importance and requiring much attention will be the collection of health data and the creation of spatial health database or databases from them, i.e., spatial health data collection and spatial health database creation (Amoly et al., 2014; Gascon et al., 2015). This could be followed by or done in parallel with the collation and creation of spatial databases for the green and blue space indicators to be used to investigate the impact of green and blue spaces on different health outcomes (Mitchell et al., 2011; Wheeler et al., 2015). This process can be described as the spatial green and blue data collation and creation. This will be followed by the selection and testing of algorithms or

methods for choosing sample sites that will be used for assessing and/or confirming relationships between green and blue spaces and human health outcomes, i.e., sample site selection (Fuller et al., 2007; Beyer et al., 2014). This will be followed by the actual assessment and confirmation of relationships between human health outcomes and the availability or non-availability of green and blue spaces using relevant statistical analysis methods, tools, or techniques (e.g., simple linear regression modelling, analysis of variance etc.) (Maas et al., 2006; Mueller et al., 2016). After confirming the relationships between human health outcomes and the availability or non-availability of green and blue spaces, other statistical techniques (e.g., principal component analysis, multiple linear regression modelling etc.) will be applied for checking the impact or contributions of other important human health confounding and/or effect-modifying factors (Lachowycz and Jones, 2013; Taghikhah et al., 2020). Furthermore, based on previously expressed uncertainties on the thresholds or limits of potential health benefits deliverable by green and blue spaces, the dynamic nature of their impacts on human health outcomes ought to be assessed using available modelling approaches or techniques, e.g., change/trend analysis, time series, spatio-temporal assessments, system dynamics etc. (Astell-Burt et al., 2014; Gascon et al., 2015; Sanders et al., 2015). Lessons from an initial year-long exploratory study applying this approach (i.e., the health-data led approach) in an Irish context are reported in the following subsections (Section 2.1-2.6). Preliminary results and discussions, as well as conclusions and recommendations on the potential wider application of the approach for other sustainability issues (e.g., health, energy, food, income, environment and climate inequality and justice issues/concerns etc.) was also discussed in the following sections (Sections 3 and 4). Conceptual diagrams showing each step of the proposed methodological framework, specific methods and indicators applied at each step of the methodological framework during the one-year explorative case study, alternative options for implementation of each step of the methodological framework, as well as likely problems to be encountered at each step of the implementation and how to address them can be found in Figs. 1–7.

2.1. Spatial health data collection and database creation

The initial step involved the collection and spatial referencing of health data types (often available only in vector format) that will be used in testing the associations between human health outcomes or conditions on the one hand and the availability and/or contact with green and blue spaces on the other hand. The exploratory study applied three vector-based datasets with different geographies at a national scale (covering the whole of the Republic of Ireland). These included secondary data on self-reported health (provided at the small area and electoral district level for 2006–2016), disability (provided at settlement level for 2006–2011) and mortality (provided at intermediate level for 2011–2014) (Foley et al., 2018a). The self-reported health and disability data were collected and made available by a national agency, the Central Statistics Office (CSO). The mortality data was initially collected and provided by another national agency, the General Register Office (GRO), but it was analyzed and re-aggregated to more workable geographies by the National Centre for Geocomputation, Maynooth University, Ireland. While other health datasets were explored, some were not made available in formats that are easily workable or convertible for further analysis (e.g., hospital records). Some others were simply not collected at a national level, hence making national scale assessment impossible (e.g., primary care data). Five major issues that were encountered within this health data collection and spatial database creation process and how they can be addressed were discussed in the following subsections. A visual description of the initial step in the methodological framework is provided in Fig. 2.

2.1.1. Data heterogeneity

It is not unusual to find the same, and sometimes multiple health data types being collected and spatially referenced using different projection systems or represented using heterogeneous and non-uniform scales or geographies. This often occurs because of poor communication on the spatial units to be used for the transmission and dissemination of public health data (Foley et al., 2018a; Pu et al., 2020). It can also occur for a variety of reasons, such as the usage of different measurement units for different health conditions, as well as the adoption of different administrative units by different health authorities/agencies concerned with reporting health data on different health conditions to the public (Rigby et al., 2017, Foley et al., 2018b). For example, within the initial, one-year exploratory case study, self-reported health data was reported at two scales namely Electoral Division (ED) and Small Area (SA) scales. Also, disability data was reported at Settlement (ST) scale; while mortality data was reported at an entirely newly derived geography referred to as Intermediate area (IA) geography or scale (Rigby et al., 2017; Foley et al., 2018a). Adopting uniform reporting units (i.e., data transmission and dissemination units) for all reported health conditions will not be possible because the health data often differ in kind and structure and can therefore not be aggregated into the same units (Foley et al., 2018b; Arodudu et al., 2019). This difficulty can however be addressed by aggregating and re-presenting all health data types available within a particular study at all scales or geographies represented within the same study i.e., health data types made available exclusively in one or two scales or geographies will be aggregated and re-presented in the other scales/geographies in which other health data types within the same study have been initially presented and vice versa, hence ensuring

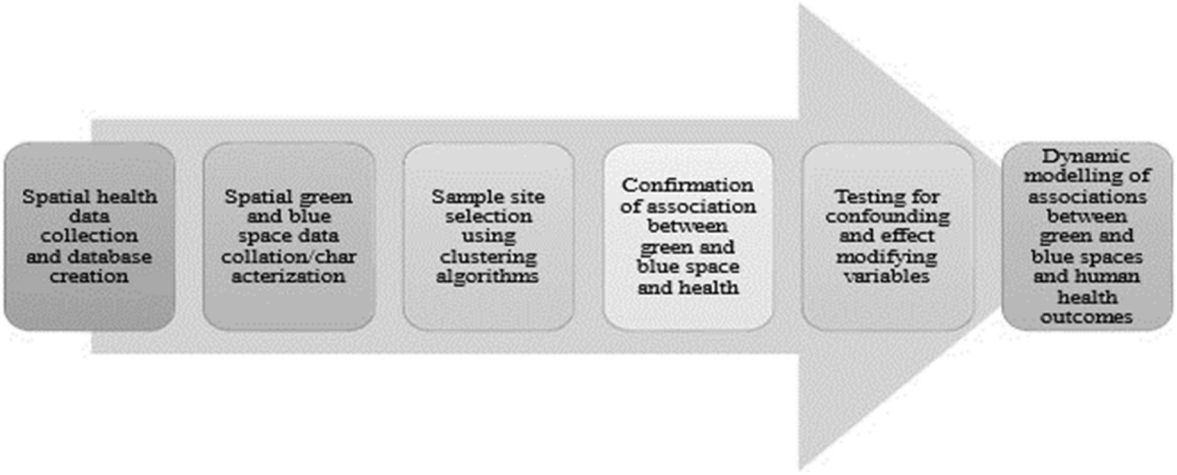

Fig. 1. A conceptual diagram demonstrating an overview of all steps involved in the proposed methodological framework.

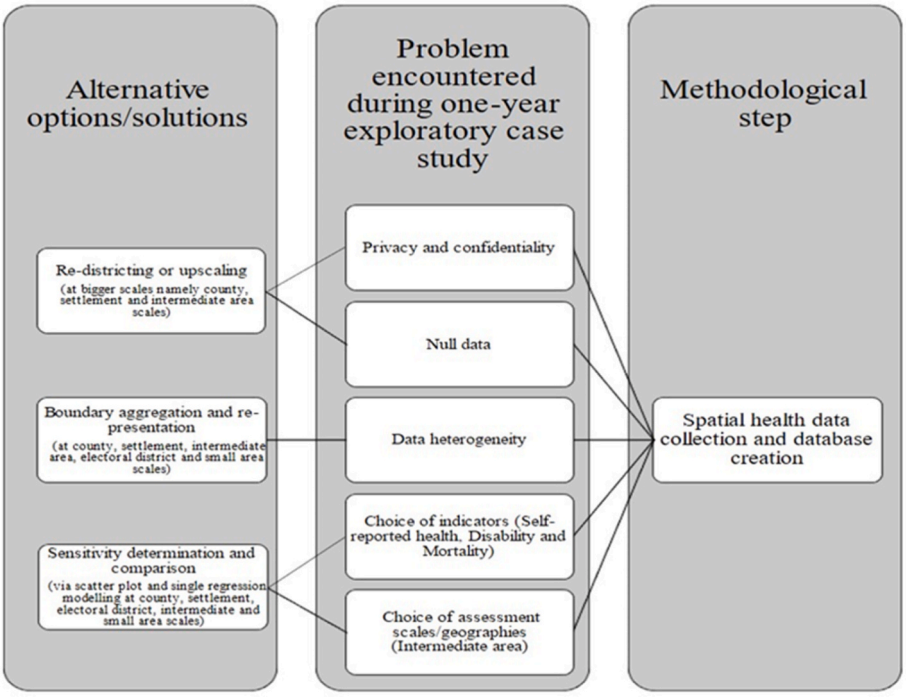

Fig. 2. Conceptual diagram showing the first step of the methodological framework (spatial health data collection and database creation), specific methods and indicators applied at the step during the one-year explorative case study, and likely problems to be encountered at the step, as well as how to address them.

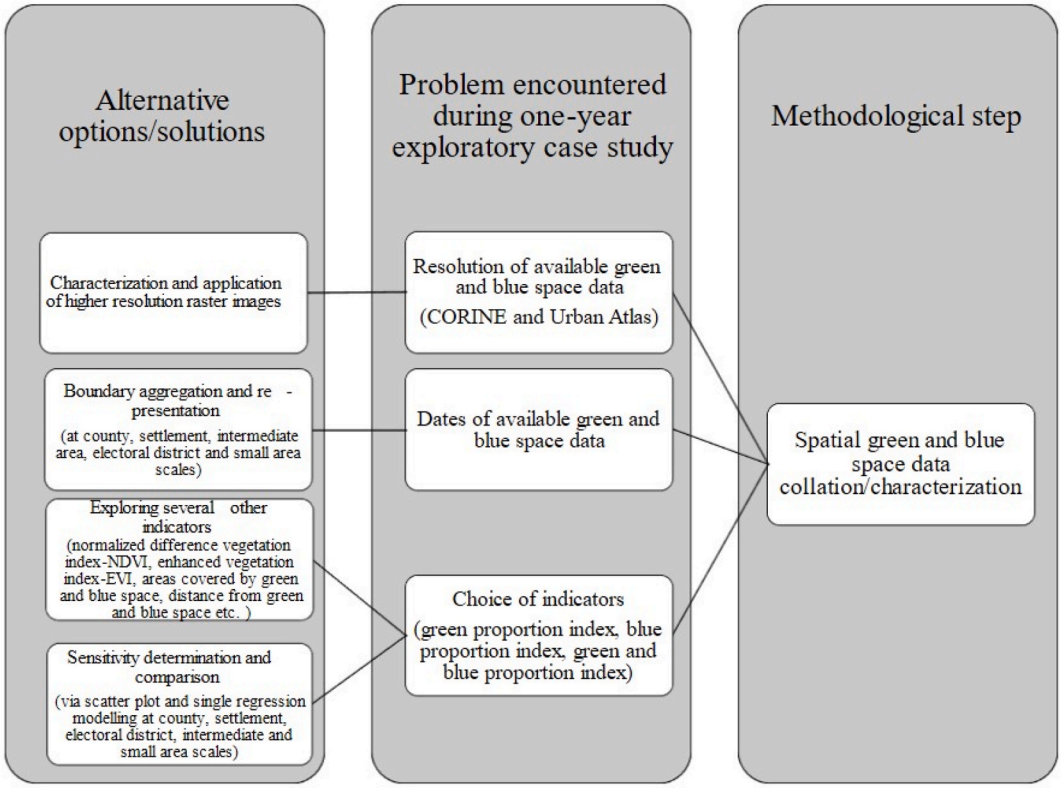

Fig. 3. Conceptual diagram showing the second step of the methodological framework (spatial green and blue space data collation/characterization), specific methods and indicators applied at the step during the one-year explorative case study, and likely problems to be encountered at the step, as well as how to address them.

uniformity and comparability across all health data types. For example, within the one-year exploration case study, self-reported health data that was initially presented at electoral division and small area scales were further aggregated and re-presented in intermediate area, county and settlement scales/geographies for uniformity and comparability. Disability data that was initially presented exclusively in settlement

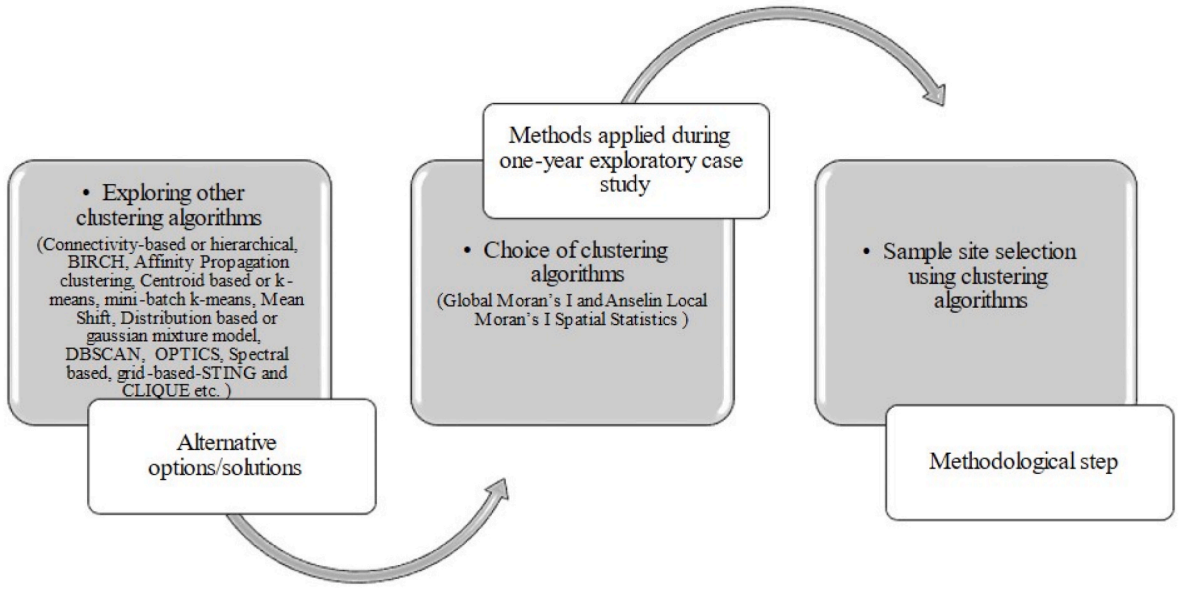

Fig. 4. Conceptual diagram showing the third step of the methodological framework (sample site selection using clustering algorithms), specific methods and indicators applied at the step during the one-year explorative case study, and alternative options for implementation of the step.

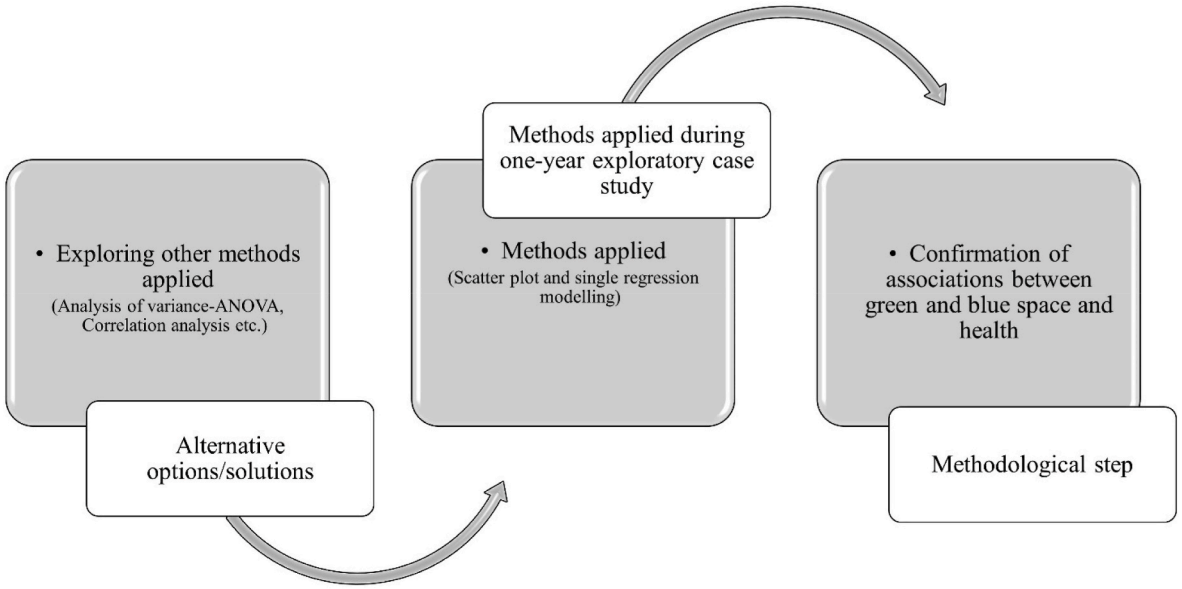

Fig. 5. Conceptual diagram showing the fourth step of the methodological framework (confirmation of associations between green and blue space and health), specific methods and indicators applied at the step during the one-year explorative case study, and alternative options for implementation of the step.

scale was further aggregated and re-presented in electoral division, small area, county and intermediate area scales/geographies for uniformity and comparability. Also, mortality data that was previously published in intermediate area geography exclusively was further aggregated and re-presented in electoral division, small area, county and settlement scales/geographies for uniformity and comparability. Boundary aggregation and re-presentation can be achieved using relevant GIS based operations, e.g., identity, union, merge, split, clip, dissolve, intersect, erase, buffer etc. One or more individual GIS operations or combinations of GIS operations may help achieve the desired boundary aggregation. What works in each boundary aggregation instance is usually determined after one or more GIS based operations must have been tested or applied once or severally.

2.1.2. Null data

Spatial representation and reporting null data for different health conditions can be challenging, especially at finer scales (Rigby et al., 2017; ESRI, 2020a). These often lead to having lots of zero or null data values associated with entire spatial data units, hence influencing the direction of further processing of the health data, i.e., selection of sample sites or spatial data units with zero values as the place with best or worst health outcomes depending on the unit or scale applied for presentation of the health conditions under consideration (Foley et al., 2018a; ESRI, 2020b). This can, by default, introduce an unintended bias into subsequent steps of the assessment process (Arodudu et al., 2019, Foley et al., 2018b). To eliminate the unintended bias that zero values can introduce into the assessment process, upscaling via redistricting (i.e., re-demarcation or merging) of smaller spatial units and delineation of larger and more viable spatial units or geographies will be required.

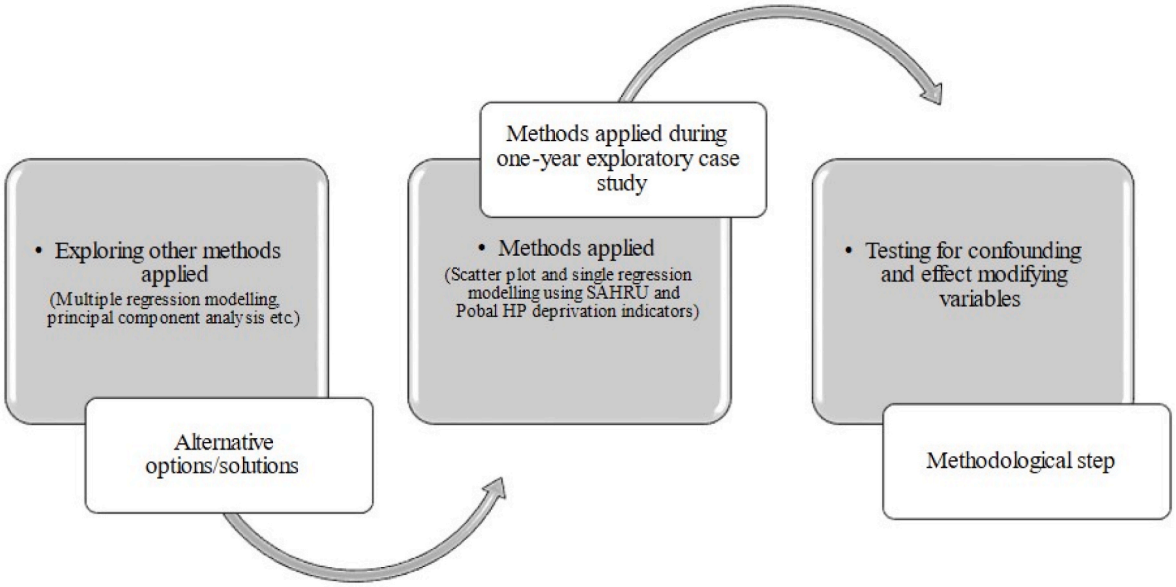

Fig. 6. Conceptual diagram showing the fifth step of the methodological framework (testing for confounding and effect-modifying variables), specific methods and indicators applied at the step during the one-year explorative case study, and alternative options for implementation of the step.

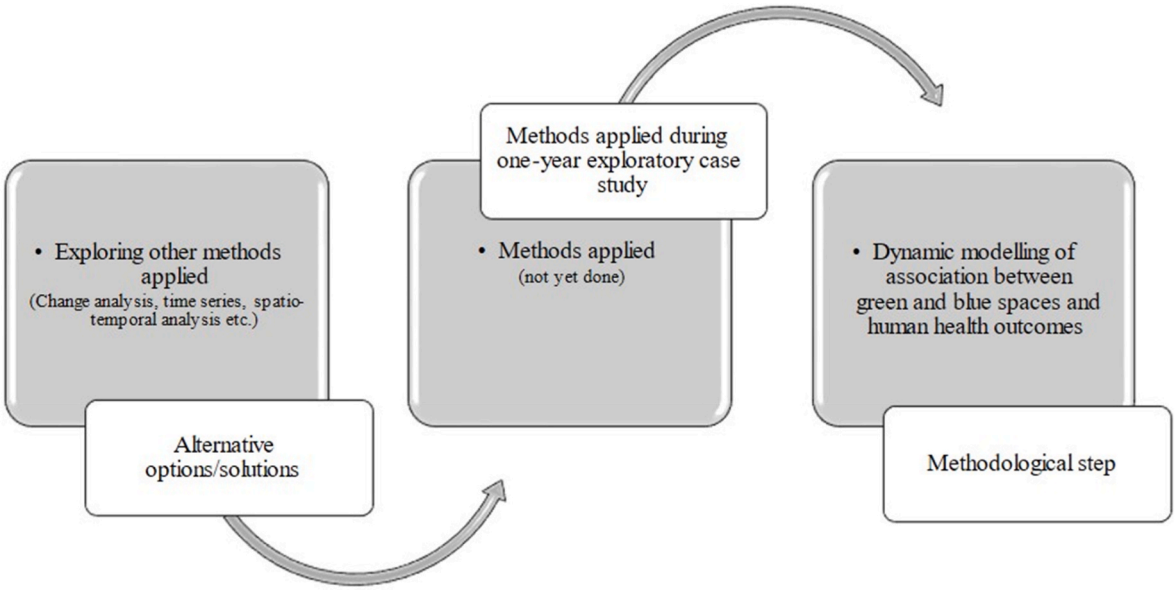

Fig. 7. Conceptual diagram showing the sixth step of the methodological framework (dynamic modelling of association between green and blue spaces and human health outcomes), and alternative options for implementation of the step.

Creation of new geographies previously unknown or unused but more fitting for the purpose of having more viable spatial units without null values, hence avoiding unintended bias or influence that zero values may cause in the process of selection of sample sites for modelling the relationships of human health outcomes, and green and blue spaces might be necessary. Redistricting can be achieved using the relevant individual as well as combinations of GIS based operations (e.g., identity, union, merge, split, clip, dissolve, intersect, erase, buffer etc.).

2.1.3. Privacy and confidentiality

National policies and laws exist in most countries protecting the privacy and confidentiality of personal data on human subjects (human health data inclusive) for the sake of safeguarding national security and the safety of persons (Foley et al., 2018b; Arodudu et al., 2018). Such laws often prohibit revealing the identity and other personal information of persons or giving clues that make the identity and other personal information (health data inclusive) of persons traceable through whichever means (Rigby et al., 2017; Foley et al., 2018a). Consequently, spatial analysis of places with good or poor health outcomes (irrespective of the health indicators applied) and the assessment of their associations with green and blue spaces should be done in a way that the identity and personal health information of persons in such areas or regions remains concealed. The analyzers and/or users of the information should not be able to trace such data back to the person(s) or groups of people concerned. This can be ensured by avoiding doing such assessments at the finest scales available or by upscaling smaller spatial units via redistricting into larger spatial units, using suitable individual or combinations of GIS based operations.

2.1.4. Choice of indicators

Several indicators can be applied to describe the same health conditions or outcomes. Different indicators for representing the same health conditions or outcomes could be quantitative or qualitative in nature or even a mix of quantitative and qualitative measures (Smith et al., 2017, Foley et al., 2018b). While some health indicators for the same health condition or outcome may be sensitive to green and blue space indicators, some may not be and vice versa (Maas et al., 2006; Arodudu et al., 2019). The best health indicators for different health conditions or outcomes are those most sensitive to associations with greenspace and blue space availability or contact indicators (Groenewegen et al., 2006; Foley et al., 2018a). To determine the best health indicators for testing the associations between a health condition or outcome and green and blue spaces, there is a need to test their levels of sensitivity and strengths of association using appropriate statistical methods e.g., regression analysis, analysis of variance, correlation analysis etc. (Lachowycz and Jones, 2013; Arodudu et al., 2018). However, the choice of indicators for this study was made based on available data (spatially referenced secondary data on self-reported health, disability, and mortality).

2.1.5. Choice of assessment scales/geographies-the need for a multi-scale/multi-geography assessment

In addressing data heterogeneity issues, the aggregation and representation of data at all scales or geographies represented in the health data types within the same study is suggested in Section 2.1.1. This solution creates a new decision point. This involves making choices on the most suitable assessment scales and geographies for confirming green and blue space and human health relationships on the one hand, and examining relationships between confounding variable indicators (e.g., socio-economic deprivation) and human health on the other hand. The choice of scales or geographies selected for spatial analysis of health outcomes (mostly in vector format), as well as confirming their relationships with green and blue spaces on the one hand, and with confounding variable indicators on the other hand can impact the outcome of the study and the improvement recommendations following on from it. The one-year exploratory case study tested the relationships between green and blue space and human health outcomes at multiple scales/geographies (county, electoral district, settlements, intermediate areas, and small areas) and found both weak and strong relationships at all scales/geographies. The study further compared the frequency and strength of associations at each scale/geography (in terms of R^2 values obtained from scatter plot and single regression modelling). The comparison revealed that the intermediate area scale/geography more frequently elicited stronger positive associations for examined green and blue space and human health outcome relationships than the four other geographies (county, electoral district, settlements, and small areas), hence its choice as scale of assessment for further confirmation of the effect of confounding variable indicators in the one-year exploratory case study (Foley et al., 2018a, Foley et al., 2018b). The use of the multi-scale or multi-geography assessment for the identification of the scale or geography at which green and blue space and human health relationships are most sensitive is therefore recommended by this study. This is because it can help facilitate an objective choice of sample sites for further confirmation of the effect of confounding variable indicators or alternative explanatory variable indicators (as the case may be). It can also help determine quantitatively the scale or geography at which the impact of green and blue space on human health outcomes is detectable/measurable, as well as at what scale or geography should health improvements be planned or implemented. Such health improvements could be in form of qualitative or quantitative landscape based green and blue infrastructure planning/interventions; or in form of socio-economic planning e.g., determining the scale of provision of jobs, health facilities, industries, and other social amenities in particular areas/regions etc.

2.2. Spatial green and blue space data collation/characterization

This next step involves examining available green and blue space geodatabases and/or creating new ones that can be described as credible or reliable enough to be applied for assessing associations between health outcomes and green and blue space. Three major challenges often encountered in doing this and how they could be addressed based on lessons from the one-year exploration study are discussed in the following subsections, namely (i) resolution of available green and blue space data, (ii) dates of available greenspace and blue data, and (iii) choice of indicators. A visual representation of the second step in the methodological framework is illustrated in Fig. 3.

2.2.1. Resolution of available green and blue space data

The resolution of available raster based green and blue space data might be a direct or indirect indicator of the accuracy and/or precision levels of the data and the data characterization process (Groenewegen et al., 2006; Arodudu et al., 2019; Bulley et al., 2023). In other words, the better the resolution of the green and blue space data the more credible or reliable might be its potential for usage in assessing associations between health outcomes and green and blue spaces (Maas et al., 2006; Arodudu et al., 2018; Ogbodo et al., 2014). Higher resolution greenspace and blue space data are mostly products of primary data (satellite imageries, aerial photographs etc.) with equally high resolutions (in terms of spatial, spectral, temporal, and radiometric) (Foley et al., 2018a; Ibrahim et al., 2022). For a more accurate evaluation of relationships between green and blue space and human health, the use of higher resolution imageries is a highly favored recommendation (Arodudu, 2013, Foley et al., 2018b). Characterization of higher resolution images for obtaining higher resolution greenspace and blue space data should also be considered if they are not too expensive for the budget size of research projects requiring such assessments. High-resolution images are mostly only commercially available, very expensive and inaccessible by low-budget exploration studies. They can, however, be obtained free from previous public or private projects that have utilized them or purchased from substantial research grants that have proposed to use them for similar or other purposes.

2.2.2. Dates of available green and blue space data

In ideal cases, green and blue space data used for examining associations between health outcomes and green and blue spaces should be of the same dates (i.e., year). It, however, often happens that available green and blue space data are not characterized for the years we have corresponding health data. Sometimes, the dates of the primary data from which green and blue space data were characterized (satellite imageries, aerial photographs etc.) are different from the date they are released publicly or published. This is because it takes lots of time to process a reliable spatial geodatabase. At other times, you often find that there is no green or blue space data for years for which health data are available. In such cases, green or blue space data of closest dates to those of available health data are assumed to be representative of what is obtainable in the actual year for which health data is available. Whenever there is a need to choose green and blue space data most representative for testing the association of near-date health outcomes with available green and blue spaces, green and blue space data of years just before those of the year of the measured or reported health data should be reasonably assumed to be more representative as they are likely to have more impact on current and near-future health outcomes than those from years afterwards. In fact, green and blue data of the current year may also not be as effective or representative in evaluating the impacts of green and blue spaces on health outcomes as those of years shortly before year the health data were measured or reported, especially if there was a drastic change in green or blue space properties and/or compositions in the actual reference year. Green and blue space data of following years also may not be as representative as those of previous years before the year the health data were measured or reported because

the green and blue space might undergo significant change in properties and/or compositions after the year the health data was measured or reported.

2.2.3. Choice of indicators

As the case is with the choice of health indicators, indicators selected to represent the potential of green and blue spaces to confer health benefits need to be seen to be the most sensitive to the health indicators whose associations are being assessed (Arodudu et al., 2018, Foley et al., 2018b). Previously, green, and blue space indicators that have been applied to assess relationships between green and blue space and health outcomes included the normalized difference vegetation index (NDVI) (Amoly et al., 2014; Beyer et al., 2014; Gascon et al., 2015), enhanced vegetation index (EVI) (Heo et al., 2020; Heo and Bell, 2023), proximity to green and blue space (Maas et al., 2006; Wheeler et al., 2012) etc. While some of these indicators (e.g., NDVI) have been applied often to confirm relationships between health outcomes and green spaces, they do not measure the role or impact of blue spaces on health outcomes i.e., NDVI classifies water bodies as zero (Beyer et al., 2014; Gascon et al., 2015). In response to this, the one-year exploration study for investigating the health-data led approach for confirming relationships between health outcomes and green and blue spaces further tested the capabilities of new green and blue space indicators to capture the potential of individuals and combinations of green and blue spaces to deliver or not deliver health benefits within the same context. The tested indicators included Green Proportion Index (GPI), Blue Proportion Index (BPI) and Green and Blue Proportion Index (GBPI). These three green and blue space indicators were measured as the proportion of green and blue spaces (individually or in combination with each other) within each boundary extent. They are usually measured between 0 and 1, with the highest value being 1 representing 100% green and/or blue space cover. The results suggest they can reveal the health benefits of greenspaces and/or blue spaces, hence could be further applied or tested alongside other previously used greenspace and blue space indicators in subsequent studies. GPI, BPI and GBPI were obtained from raster-based CORINE (CO-ordination of infoRmatIon on the enviroNmEnt) land cover information for 2006 and 2012 (with a spatial resolution of 30 m) and Urban Atlas (UA) land cover data (with a spatial resolution of 2.5 m). While urban atlas land cover data has better spatial resolution, it is only available for limited areas (mostly large urban centers in Ireland and their adjoining towns and counties namely Cork, Limerick, Dublin, Waterford, and Galway). CORINE land cover data on the other hand is available for the whole of Ireland but with less spatial resolution. Urban Atlas land cover data has more well-defined green space land cover classes, while CORINE has more well-defined blue space land cover classes.

2.3. Sample site selection using clustering algorithms

After collection, collation, creation/characterization of spatial health data and green and blue space data, the next step or process within the health data led approach involves choosing samples sites of good and poor health outcomes most suitable for testing associations between health outcomes and green and blue spaces. From the initial one-year explorative study, we inferred that data clustering methods can play a role in facilitating this objective. Since the identification of sample points of good and poor health outcomes is central to the health data led approach, we deduced that examining the potential role of data clustering methods for identifying and generating clusters of sample points of places with good and poor health outcomes for subsequent investigation of relationships between health outcomes and green and blue space could offer some new insights.

During the one-year explorative study, we tested the Global Moran's I Spatial Statistics/Clustering method and Anselin Local Moran's I Spatial Statistics/Clustering method in ArcGIS software. Other data clustering methods that can also be applied include connectivity-based or hierarchical clustering (Agglomerative or bottom-up hierarchical clustering and Divisive or top-down hierarchical clustering), BIRCH (Balance Iterative Reducing and Clustering using Hierarchies), Affinity Propagation clustering, Centroid based clustering or k-means algorithm, mini-batch k-means, Mean Shift, Distribution based clustering (Gaussian mixture model), Density based clustering (DBSCAN-Density-based spatial clustering of applications with noise and OPTICS-Ordering points to identify the clustering structure), Spectral based clustering, grid based clustering (STING and CLIQUE) etc. The different data clustering methods mentioned have different underlying principles and statistical underpinnings that they can employ in generating clusters of sample points of places with good and poor health outcomes for subsequent assessment of associations between health outcomes and green and blue spaces. This will be discussed in the following subsections. A conceptual illustration of the third step in the methodological framework is depicted in Fig. 4.

2.3.1. Global Moran's I spatial statistics

This clustering algorithm generates spatial data clusters from spatial health data, as well as other kinds of spatial data describing phenomena (Anselin, 1995; Getis, 2010). It is however susceptible to spatial autocorrelation, i.e., it also adds neighboring features that are close in attribute values to the generated clusters, as it considers them a product of or part of the generated cluster without prior checks (Anselin, 2005; Grieve, 2011). With regards to the health data led approach, the Global Moran's I will identify places with good and poor health outcomes and their spatially autocorrelated surrounding areas, hence generating clusters based on the identified information.

2.3.2. Anselin Local Moran's I spatial statistics

This clustering algorithm identifies statistically significant hot spots, cold spots and spatial outliers in spatial health data and other kinds of spatial data (Anselin, 2005; Helbich et al., 2012). Unlike the Global Moran's I, which measures and considers spatial autocorrelation in determination of clusters, it applies False Discovery Rate (FDR) Correction which removes all spatially autocorrelated values leaving only statistically significant clusters and outliers (with 95% confidence level) (Li et al., 2007; Alvioli et al., 2016). With regards to the health, data led approach, it identifies places with good and poor health outcomes and generates a cluster from the information having filtered out surrounding spatially autocorrelated areas. It usually forms two types of cluster-LL (statistically significant cluster of low values) and HH (statistically significant cluster of high values), as well as two types of outliers LH (feature with low values surrounded by features with high values) and HL (feature with high values surrounded by features with low values) (Anselin, 2005; Getis, 2010).

2.3.3. Connectivity-based or hierarchical clustering (agglomerative or bottom-up hierarchical clustering and divisive or top-down hierarchical clustering)

Connectivity-based, also known as hierarchical clustering, identifies the most similar objects as a cluster assuming that the neighbor objects are more related together in comparison to the other objects that are further away. The agglomerative clustering algorithm is the most popular method in this category. This method groups health data in a set of clusters based on differentiations and similarities among provinces' distances. It treats each object in the data as a potential cluster and then moves up the hierarchy while merging pairs of clusters together (Delil et al., 2017; Kassambara, 2019). Another kind of algorithm for hierarchical clustering is divisive clustering. Unlike the previous method, here, clustering is progressed from top to bottom of the hierarchy. It considers the whole objects in the given dataset as a single cluster, and then recursively divides into multiple clusters from the root to the last pieces (Zhang et al., 2017). The disadvantage of employing hierarchical cluster analysis is its scalability. Clustering a large size of health data observation with this method would be highly computationally expensive

(Dunning, 2020).

2.3.4. Balance Iterative Reducing and Clustering using hierarchies

The Balance Iterative Reducing and Clustering using Hierarchies (BIRCH) method is mainly used in clustering big data (Zhang et al., 1996; Han et al., 2022). When the algorithm reads the input data, a compact version of the dataset is created, which is a summary of the original dataset containing as much information as possible (Zhang et al., 1995; Ramadhani et al., 2019). Then, instead of clustering the bigger dataset, the smaller summary dataset is clustered (Zhang et al., 1997; Chiu et al., 2001). BIRCH is commonly used in integration with other clustering algorithms (Fichtenberger et al., 2013; Lang and Schubert, 2020). While it can rescale the dataset to a more affordable data size by summarizing, other methods can be utilized to enrich the clustering section performance (Zhang et al., 1999; Lang and Schubert, 2022). In health research, it has been used for patient profiling and disease detection (Alashwal et al., 2019; Razzak et al., 2020; Shuai, 2022). The major limitation of this algorithm is that it can only work with metric attributes, for example, vector values produced in a Euclidean distance, and it is unable to work with categorical attributes (Charest and Plante, 2014; Gupta, 2021).

2.3.5. Affinity Propagation clustering

The Affinity Propagation (AP) clustering algorithm was firstly developed by Frey & Dueck (2007). In this method, each cluster is randomly assigned an “exemplar” or sample. These data point subsets are iteratively refined until they end up with the best choice by comparing similarities via AP clustering which identifies the sample that best represents the cluster. Similar to the hierarchical clustering approach, it can categorize the objects by determining the similarity among them. Moreover, AP addresses the limitation of BIRCH and many other prototype-based clustering methods, such as K-means, as it is not limited to metrics attributes only (e.g., vector space structure). This method has application in clustering genetic codes (Jianjun and Jianquan, 2018) and identifying functional networks of the brain (Zhang et al., 2015). AP can be a suitable approach for health data clustering since many similarity measurements that are applied in health applications are not directly related to a clear vectorial description (Bodenhofer et al., 2011).

2.3.6. Centroid-based clustering (k-means algorithm)

Centroid-based clustering is a method of clustering spatial health data into non-hierarchical clusters. One of the most applied algorithms of this method is k-means, where the given database with ‘n’ objects is split into a set of “K" clusters. This unsupervised machine learning algorithm starts with a randomly selected first set of centroids as an initial point of each cluster. Afterwards, it iteratively optimizes the position of the selected center points. This optimization halts when either the centroids have stabilized, and no more changes occur, or the defined number of iterations has been reached (Santhanam and Padmavathi, 2014). This algorithm has limited application in health research but has been recently used in detecting subpopulations of cells (Wardani et al., 2019).

2.3.7. Mini-batch k-means

As the name indicates, the Mini-Batch k-Means is a version of the standard k-means algorithm developed for big data. Instead of iterating over the entire dataset, it works with random batches to reduce stochastic noise and computational costs (Sculley, 2010). It selects a fixed number of data randomly and stores them in small size batches in the memory. Then, after selecting a random sample of data, it iteratively updates the batches and reduces the convergence time. In the health context, for example, gene data can be clustered quickly in a scalable and memory-efficient manner (Hicks et al., 2021). One limitation of this algorithm is that it can be easily trapped in the local optimal solutions, which can negatively impact the performance of clusters (Xiao et al., 2018).

2.3.8. Mean shift

Mean shifts, also known as mode seeking algorithm, is a centroid-based algorithm that belongs to the unsupervised learning group of algorithms. It has many applications in image processing and computer vision. In contrast to the K-Means clustering algorithm, the number of clusters in this algorithm does not need to be pre-decided and will be automatically specified with respect to the data. It shifts data points towards the mode (i.e., cluster centroids are the highest density of datapoints in the region) and iteratively assigns the data points to the clusters. For any given dataset, this algorithm estimates the underlying probability density function by adding the individual kernels (e.g., Gaussian) (Tripathi, 2022). With regard to health research, mean shifts are widely used in mining medical images to extract association rules and hidden information (Cui et al., 2022). Its main limitation is related to expensive computational costs.

2.3.9. Distribution-based clustering (Gaussian mixture models)

Distribution-based algorithms group data points according to their likelihood of belonging to the same probability distribution. Instead of proximity (similarity/distance) and composition (density) in other algorithms, they consider probability as the clustering metric. The probability of getting included in a cluster is higher for those data points that are closer to the cluster center. For example, the gaussian mixture model (GMM) is a method of such that assumes that the data have Gaussian distributions. Through an interactive optimization process for fitting data, it allocates data points to the K number of clusters, which their mean, covariance and mixing probability are estimated using the Expectation Maximization technique. Distribution-based algorithms have wide applications in patient and disease phenotype clustering to understand disease pathophysiology, predict treatment response (Alhasoun et al., 2018; Loftus et al., 2022). The advantage of this algorithm is that it is flexible, and the cluster shape does not need to be defined. However, if most data points do not belong to a predefined distribution, the algorithm can easily get trapped in an overfitting issue (Joshi, 2022).

2.3.10. Density-based clustering (DBSCAN and OPTICS)

Density-based clustering method uses areas of high and low data points concentration to form clusters that vary in shape and size. The extracted clusters have the highest degree of homogeneity as the noise and outliers are excluded (Kriegel et al., 2011). This method does not require a prior specification of the number of clusters. DBSCAN (Density-based spatial clustering of applications with noise) and OPTICS (Ordering points to identify the clustering structure) are the two main algorithms of this method. DBSCAN groups contiguous regions in data that have the high density to form clusters and use low density points to distinguish them from the others. By determining the minimum number of data objects in a pre-selected radius, the algorithm determines whether the data points fit the cluster. It does the clustering operation by calculating the density reachability and density connectivity measures (Campello et al., 2020). In the DBSCAN method, only one set of hyperparameters (known as global parameters) can be selected to do the clustering for different densities. In addressing this issue, OPTICS was developed. Instead of explicitly producing a clustering data set, it calculates an augmented cluster ordering that includes a wide range of parameter settings to prioritize objects with higher density. In addition to ordering, for each object, two values of core-distance and a suitable reachability-distance will be stored and used in extracting clusters (Ankerst et al., 1999). Mining medical images (Celebi et al., 2005), phenotype clustering (Loftus et al., 2022), and medical diagnosis (Waqas et al., 2022) are only a few examples of density-based clustering method applications in the health domain. The limitations of this method are related to its low-performance in clustering high-dimensional data (Taghikhah et al., 2021) and its failure to deal with neck type datasets

with high variabilities of densities (Moreira and Santos, 2005).

2.3.11. Grid-based clustering (STING and CLIQUE)

The grid-based clustering method clusters data points using a multi-resolution grid data structure and grid cells. This structure is made up of a predetermined number of cells containing the object areas. STING (A Statistical Information Grid Approach) (Wang et al., 1997) and CLIQUE (Clustering In QUEst) (Agrawal et al., 1998, 2005) are two interesting algorithms of this method. The STING divides the spatial area into rectangular cells that are multi-level. Using a top-down approach, it iteratively partitions the high-level layer into several smaller cells in the next lower level. The higher level uses low-level statistics and parameter values for estimating its parameters and responding to queries. This algorithm is easy to parallelize and can be incrementally updated; however, it fails to detect any diagonal boundary for the clusters (DM365, 2020). CLIQUE is a combination of density- and grid-based methods that automatically identifies the high dimensional data sub-spaces to improve the clustering of the original space. It divides the dimensions into rectangular shaped units that have equal-length intervals and do not overlap and then connects the dense units to form a cluster within a subspace. This algorithm uses the size of input to scale linearly and can efficiently deal with scalability issues with the increasing number of dimensions in the data. Nevertheless, as its simplicity increases, the clustering accuracy drops. In health research, these algorithms are used for handling noisy data when clustering health behavior (Rabel et al., 2019) and detecting disease clusters (Neill and Moore, 2004). The main advantage of the grid-based clustering method is its short processing time. Irrespective of the number of data objects, the number of cells in each dimension has the greatest impact on its performance.

2.3.12. Spectral clustering

Spectral Clustering is an efficient algorithm for converting complex multidimensional datasets into low dimensional space through transforming the clustering problem into a graph-partitioning problem (Thanniru, 2022). Rooted in graph theory, this algorithm treats each data point as a graph and considers the edge connecting nodes to identify communities of nodes and segregate them for extracting clusters. It is a flexible, relatively fast (for small datasets), and easy-to-implement algorithm that makes no assumptions about the cluster forms and can be used for non-graphical data as well. With regards to applications in the health area, it is a powerful and versatile technique for medical image analysis (Schultz and Kindlmann, 2013), patient profiling (Pellicer-Valero et al., 2020), and care prioritization (Roman, 2021). One limitation of spectral clustering is related to its assumption about spherical position of data points in a cluster around its center, which might not be always relevant. The other shortcoming is that for large datasets, this algorithm is computationally expensive and inaccurate since computing the convex boundaries increases the complexity significantly (Thanniru, 2022).

Many of the data clustering methods described in this paper have only been experimented with micro-level health data e.g. For patient profiling, disease detection, medical image analysis, care prioritization, health behavior observation, mining medical images, phenotype and gene analysis, disease pathophysiology, treatment response prediction, functional brain network identification, cell subpopulation detection etc. Noteworthy however is the fact that most studies evaluating the impact of green and blue space on human health are done at macro-level. There is therefore the need to test many of the data clustering methods suggested for application within the context of a health data led approach by this study for suitability at the macro-level. Nevertheless, the inherent spatial autocorrelation effect of the Global Moran's I method and the removal of the spatial autocorrelation effect in the Anselin Local Moran's I method, both methods were still found suitable for choosing places with good and poor health outcomes at macro-level from spatial health data, hence applicable for a health data led approach during the one-year exploratory study. There is however the need to still ascertain if the other clustering methods that have only been applied on micro-level health data and still have limited spatial health data applications can do the same.

2.4. Confirmation of associations between green and blue spaces and human health outcomes

After choosing sample sites for assessing associations between health outcomes and green and blue spaces, the next step will be to empirically assess and confirm associations between them using appropriate statistical analysis methods e.g., single regression modelling, analysis of variance, correlation analysis etc. While most studies of this nature randomly apply these statistical methods, there needs to be improved understandings on when and why we use these different methods i.e., in what circumstance is one method more appropriate or valuable than the other (Lachowycz and Jones, 2013; Foley et al., 2018a). The motivation of the statistical method adopted by every study for confirming associations between health outcomes and green and blue spaces is useful for reproducibility of studies of this nature. This exploratory study only assessed the individual relationships between green and blue indicators (independent variables) and good and poor health indicators (dependent variables) hence the use of scatter plots and single regression modelling. A visual illustration of this fourth step of the methodological framework can be found in Fig. 5.

2.5. Testing for confounding and effect-modifying variables

After an initial confirmation of relationships between health outcomes and greenspaces/blue spaces, there is also a need to check for the potential impacts of confounding (co-determinants of health outcomes) and effect-modifying variables (not co-determinants but affects co-determinants hence affects health outcomes) on health outcomes. Examples of such variables may include gender, age, race/ethnicity, climate/climate change, policy, income level, socio-economic deprivation etc. Some of these variables may have more effects on health outcomes than presence, absence or contact with green and blue spaces while some have less, this needs to be ascertained in each study or context. In some cases, green and blue spaces may even be confounding or effect-modifying variables. Statistical methods that could be deployed for this include principal component analysis, multiple regression modelling etc. Due to the exploratory nature of this study, we applied scatter plots and single regression modelling as only individual relationships between the confounding variable (i.e., socio-economic deprivation) and health outcomes were tested and not joint relationships. A diagrammatic representation of this fifth step of the methodological framework can be found in Fig. 6.

2.6. Dynamic modelling of association between green and blue spaces and human health outcomes

Dynamic modelling methods are developed to analyze the time dimension and sequence of actions in physical and non-physical systems. After checking for confounding and effect-modifying variables, there is a need to also ascertain if the impacts of green and blue spaces on health outcomes have been static or dynamic. To do this, we can conduct change analysis, time series or spatio-temporal analysis to see if over time, the relationship between green and blue spaces and health outcomes has been dormant, active, continuous, or variable. Time series analysis is "an ordered sequence of values of a variable at equally spaced time intervals" (NIST, 2012). This method analyses the mechanism underlying the dynamics of observed patterns and provides feedback and feedforward predictions. Spatio-temporal analysis is an emerging field that considers both space and time dimensions of a phenomenon to extract the trajectories. In whatever form (i.e., either as change analysis, time series or spatio-temporal analysis), facilitating the dynamic

modelling of associations between health outcomes and green and blue spaces will be highly dependent on the availability of continuous and representative high-resolution green and blue space data, as well as spatial health data of as many years as possible. A graphical description of this sixth step of the methodological framework can be found in Fig. 7.

3. Preliminary results and discussions

Preliminary results from the one-year exploratory study applying the health data led approach (using national health data of Ireland-self-reported health data, mortality data and disability data) suggests higher probability of revealing both weak and strong relationships between green and blue space indicators and health outcome indicators, as well as between confounding variables and health outcome indicators due to the objectivity associated with using clustering algorithms to arrive at the sample site choices (Arodudu et al., 2017; Foley et al., 2018a). These results are explained in (i)-(iii). While previous approaches only seek to show detected positive relationships between green and blue space, and human health indicators, the health data led approach adopted by this study applied two clustering algorithms (Global Moran's I and Anselin Local Moran's I Spatial Statistics/Clustering algorithms) for sample selection and revealed both weak and strong relationships between green and blue space indicators and health outcome indicators, as well as between confounding variables (i.e. socio-economic deprivation) and health outcome indicators both successively (i.e., one after the other-confirming relationship between green and blue space and health outcomes first, then followed by confirmation of relationships between socio-economic deprivation and health outcomes), as well as in parallel to each other (i.e., evaluating the relationships between green and blue space and health outcomes, and that between socio-economic deprivation and health outcomes at the same time). Two indices derived from an Irish context for measuring socio-economic deprivation were tested as indicators for the confounding variable (i.e., socio-economic deprivation). This includes the SAHRU- Small Area Health Research Unit index and the Pobal HP (Haase and Pratschke) deprivation index. Similarly, the results of testing socio-economic deprivation as a confounding or effect-modifying variable also suggests both weak and strong relationships between socio-economic deprivation indicators and good and poor health outcomes under different circumstances (Arodudu et al., 2017; Foley et al., 2018a). The different shades of preliminary results obtained from the one-year exploratory study are described and presented in the following subsections i-ii; and Figs. 8–10.

(i)Contrary to assumptions that higher proportion of green and blue spaces, and high level of affluence (i.e., low socio-economic deprivation) should result in better health outcomes, Fig. 8a suggests conversely that there could be close associations between places with high socio-economic affluence or low socio-economic deprivation (represented by SAHRU deprivation index) and long-term disability (represented by % long-term disability); Fig. 8b suggests that places with green spaces (represented by green proportion index) may still have relatively high self-reported poor health (represented by Kavanagh-Foley index of wellbeing); while Fig. 8c suggests that in certain circumstances, there might be closer associations between places with blue spaces (represented by blue proportion index) and self-reported poor health (represented by 3-point self-reported health data), especially when there are no frequent contacts between green and blue spaces and people in their neighborhoods.

(ii)On the other hand, while Fig. 9a suggests that there might be a positive close association between high socio-economic affluence or low socio-economic deprivation (represented by SAHRU deprivation index) and low mortality rates (represented by life expectancy under the age of 75), Fig. 9b and 8a suggests that there could still be a negative relationship between high socio-economic affluence or low socio-economic deprivation (represented by another deprivation index-Pobal HP deprivation index) and another health indicator (long-term disability as

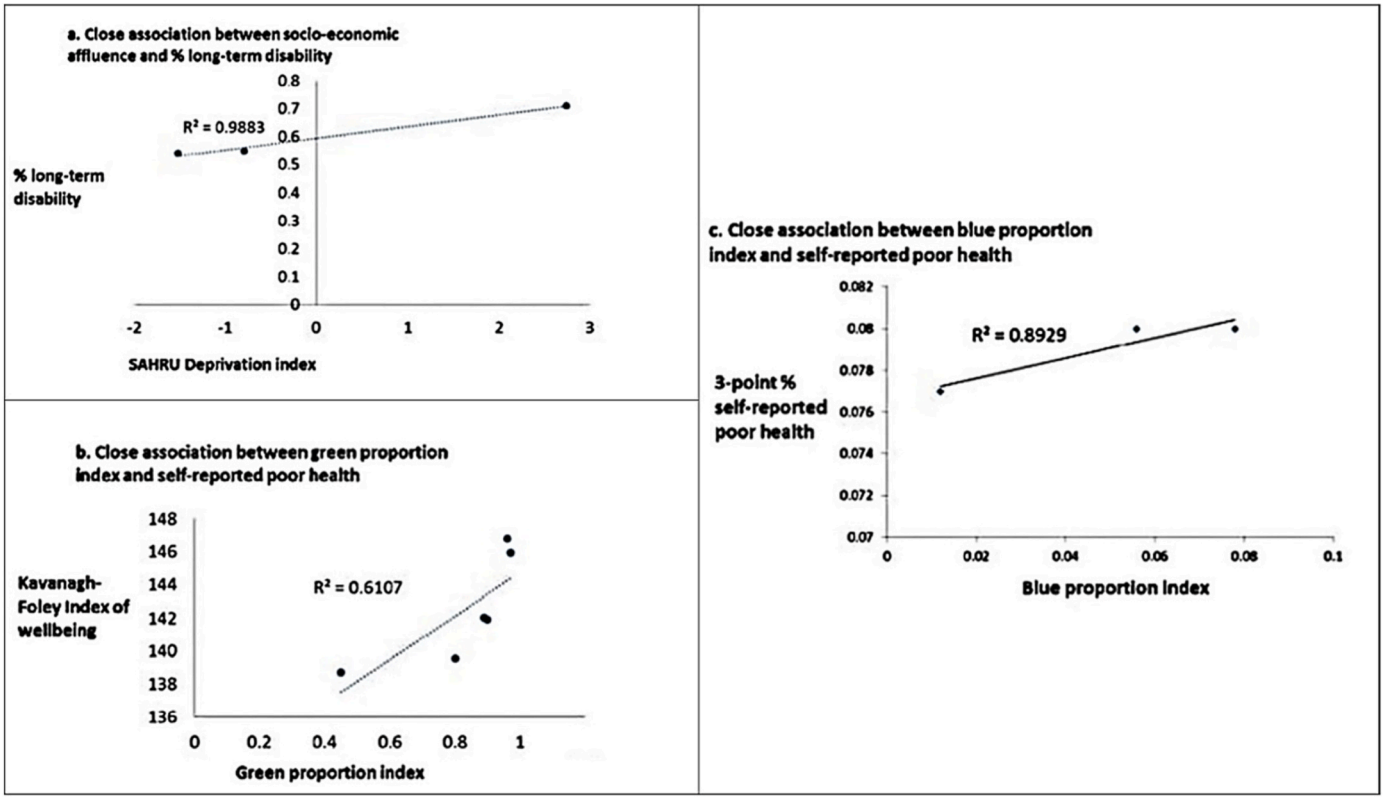

Fig. 8. Different relationships observed between green and blue spaces and human health, as well as between socio-economic deprivation/affluence and human health (Foley et al., 2018a).

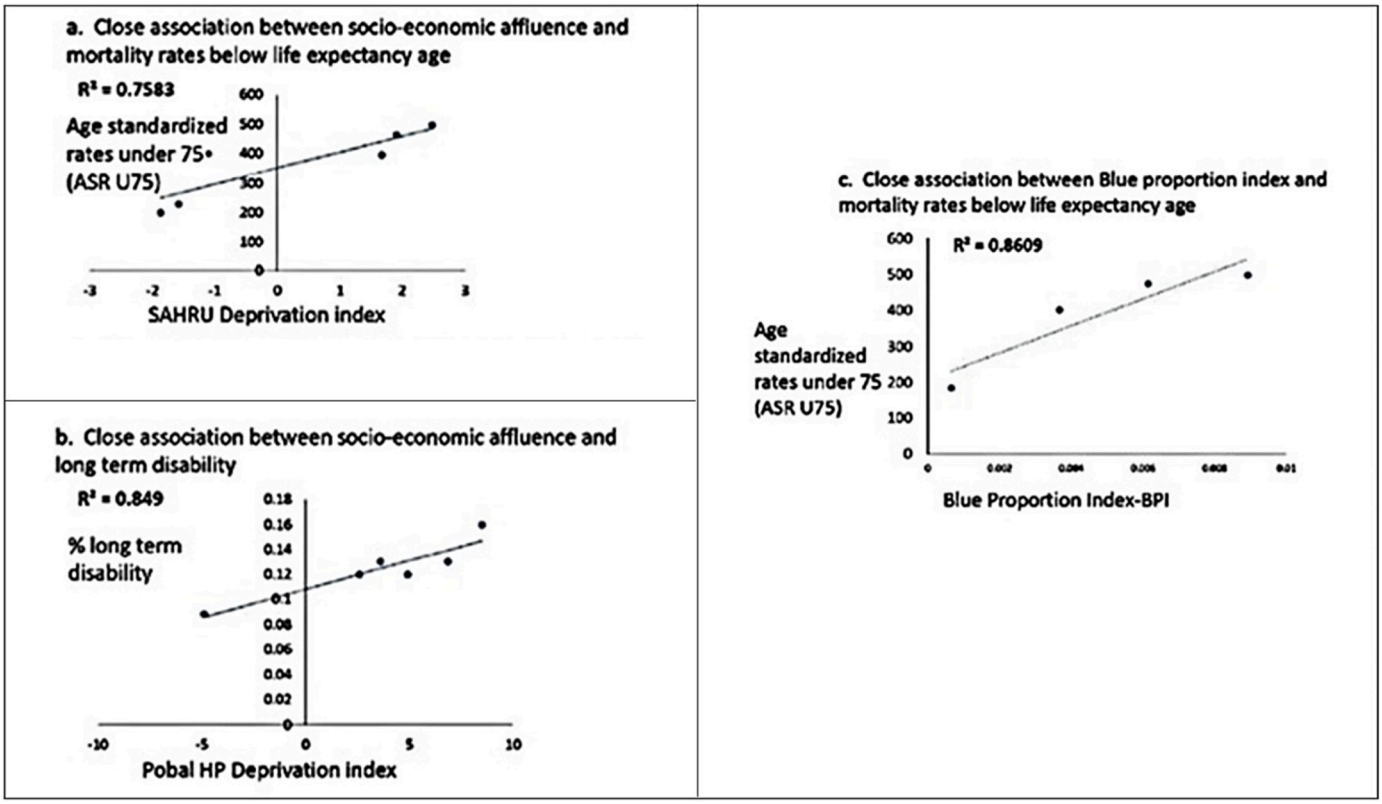

Fig. 9. Comparatively different relationships observed between green and blue spaces and human health, as well as between socio-economic deprivation/affluence and human health (Foley et al., 2018a).

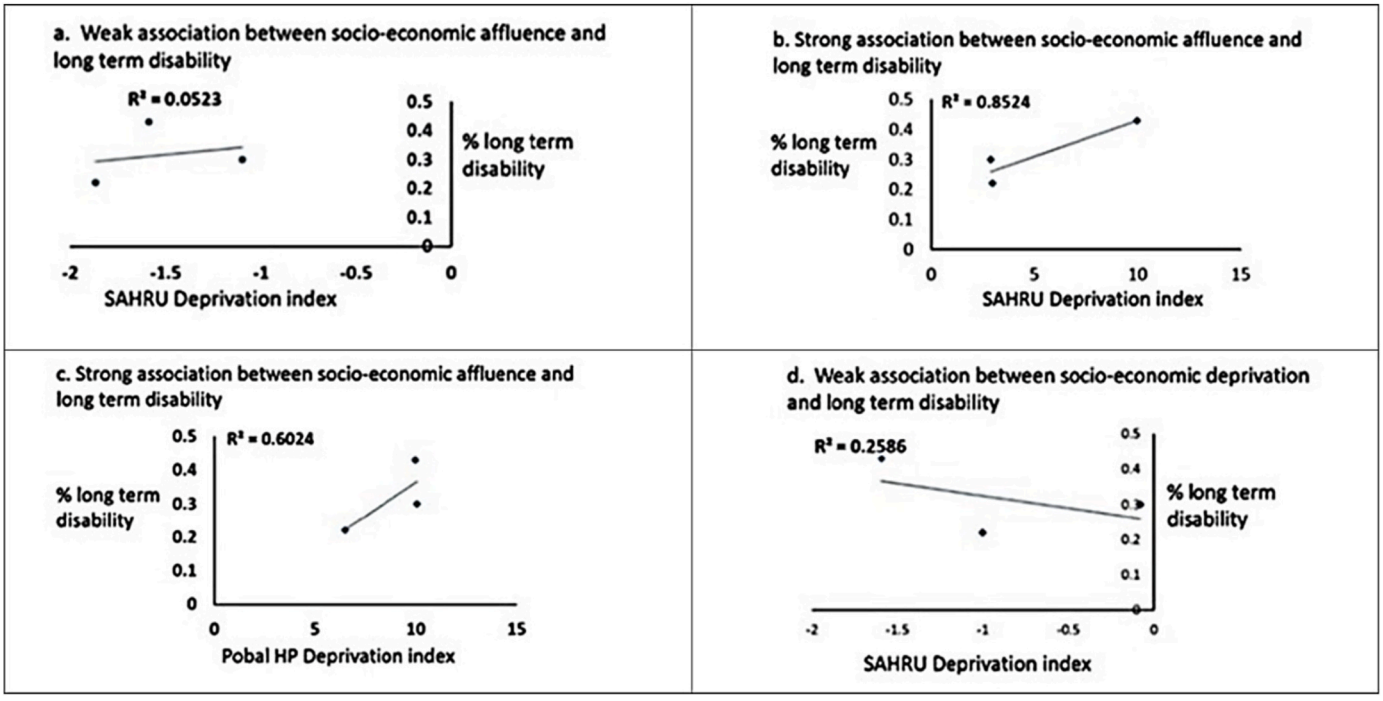

Fig. 10. Conflicting associations between socio-economic deprivation/affluence and health (Foley et al., 2018a).

represented by % long-term disability). In contrast to Fig. 8c which reveals close associations between blue space availability (represented by blue proportion index) and poor health indicator (represented by 3-point self-reported health data), Fig. 9c on the other hand confirms close associations between blue space availability (represented by blue proportion index) and low mortality rates (represented by life expectancy under the age of 75). Hence the difference in the impact of blue spaces in Fig. 8c and 9c may have been the level of contact between blue

spaces and the people living in their surrounding neighborhoods. This was however not further investigated within the framework of this one-year exploratory study.

(iii)More specifically and in contrast to results in Fig. 8a, 9b and Fig. 10a found weak associations between high socio-economic affluence or low socio-economic deprivation (represented by SAHRU deprivation index) and long-term disability (represented by % long-term disability) at a different instance. This also contrast with results in Fig. 10b and c (but aligns with Fig. 8a and 9b) which indicates strong associations between high socio-economic affluence or low socio-economic deprivation and long-term disability after testing the relationships of two different socio-economic deprivation indices (SAHRU and Pobal HP deprivation indices) against the same health indicator (represented by % long-term disability) at different instances. In reality, the strong association of people with high socio-economic affluence or low socio-economic deprivation with long-term disability in these instances (Fig. 8a, 9b and 10b and 10c) might be as a result of deliberate government residency plan/policy, location or settlement of people with long-term disability (e.g., old people, retirees, refugees from war-torn countries etc.) in the same areas of jurisdiction with people of higher socio-economic influence in the society, hence the observed pattern. Fig. 10d (in contrast to Fig. 9a) also found weak associations between socio-economic deprivation (represented by SAHRU deprivation index) and long term-disability (represented by % long-term disability).

Readers can find a full set of complete results of the study as presented to and subsequently published by the funders, the Environmental Protection Agency of Ireland (EPA) in its EPA Research Report No. 264. https://www.epa.ie/publications/research/environment–health/Research_Report_264.pdf(epa.ie).

4. Conclusions and recommendations

This paper synthesized lessons learnt from the one-year exploration study applying the health data-led approach and shows a few results as a proof-of-the-concept. The methodology and results of the one-year case study are documented by Foley et al. (2018a). The Global Moran's I Spatial Statistics/Clustering algorithm, and the Anselin Local Moran's I Spatial Statistics/Clustering algorithm (both from the ArcGIS software) proved valuable tools in identifying places with good and poor health outcomes, as well as assessing relationships between green and blue spaces and health outcomes. Unlike previous studies that only seek to reveal detected positive relationships between green and blue space, and human health indicators, the objectivity associated with using the two clustering algorithm (namely Global Moran's I and Anselin Local Moran's I Spatial Statistics/Clustering algorithm) helped to reveal both weak and strong relationships between green and blue space indicators and health outcome indicators, as well as between confounding variables (i.e., socio-economic deprivation) and health outcome indicators using only scatter plots and single regression modelling methods. The choice of sample sites/locations by the algorithms can be a pointer to the potential causes or determinants of good and poor health because values of the independent variables (i.e., green and blue space indicators and confounder variables) can easily be associated with values of dependent variables (indicators of good and poor health outcomes) even before extensive confirmation of relationships. This is the reason, confirmation of relationships between green and blue space indicator and health outcome indicators, as well as between confounder variable indicator (i.e., socio-economic deprivation) and health outcome indicators under this exploratory study was done using only scatter plots and single regression modelling methods. At the initial exploratory stage of the analysis, the scatter plot and single regression modelling methods already revealed both weak and strong relationships between health outcomes and green and blue space indicators on the one hand, and between health outcomes and confounding variable indicators (i.e., socio-economic deprivation) on the other hand. The study therefore did not need an extra procedure to confirm the presence or absence of confounding variables as one had been identified already (socio-economic deprivation). That said, the use of multiple regression modeling, principal component analysis and other more advanced techniques will still be needed to further quantify and/or compare the impact or role of green and blue space indicators to those of other confounding variable indicators in determining good or poor health outcomes. Due to the effectiveness of the two clustering algorithms/methods deployed by this study, the testing of other machine and/or deep learning clustering algorithms and methods (e.g., connectivity-based, or hierarchical, BIRCH, Affinity Propagation, Centroid based, Distribution based, DBSCAN, OPTICS, STING and CLIQUE clustering methods) is highly recommended by this study to determine and compare their suitability for this same purpose.

Also noteworthy is the fact that data-led or data-first principles and procedures recommended by this paper, which utilizes observed/characterized data and data clustering algorithms/methods for choosing sample sites for investigation of relationships, is not only valuable for assessing/confirming associations between good/poor health outcomes and green and blue spaces. It can also be applied for identifying places with good or poor conditions of living (e.g., in terms of food access, energy access, sustainable income, poverty levels, shelter access etc.), before confirming the impact of race, ethnicity, institutional exclusion, gentrification etc. In driving such living conditions/outcomes. This is because previous research of this nature often associates poor conditions of living with race, ethnicity, institutional exclusion, gentrification etc. by default (Bullard, 2001; Gardner-Frolick et al., 2022). Such studies need an initial evaluation and/characterization of living conditions to prove and substantiate claims of poor living conditions (Wing, 2005; Gonzales, 2022). This will help provide better intelligence information for environmental planning, especially with regards to prioritizing areas most badly affected by poor living conditions to ensure environmental justice and equity (Bullard, 2003; Martin, 2021). This is important for distribution of climate mitigation/adaptation projects, greenspace, food, renewable energy etc. A data-led or data-first approach or paradigm can also assist in answering specific space and time related questions within the context of sustainability assessments. Previous sustainability assessment often concentrates on assessing the social, economic, and environmental impacts of products, processes, projects, plans and policies in-situ and at mid and/or end points without measuring the extent of problems that has been created over space and time (Martuzzi et al., 2010; Menton et al., 2020; Taghikhah et al., 2022a). Within sustainability assessment contexts, the data-led or data-first paradigm suggested by this study can help confirm relationships between sustainability challenges and their driving or causative factors over space and time on the one hand, while also more accurately assist in defining the local, regional or national sustainability aspirations and/or sustainable development goals towards remedying them on the other hand (Song et al., 2020; Calderón-Argelich et al., 2021; Taghikhah et al., 2022b). The data-led or data-first paradigm is therefore valuable, transferable, effective, and applicable for providing relevant information for solving different sustainability problems, especially within the emerging fields of sustainability assessments/analysis, environmental sustainability, climate change mitigation and adaptation, climate justice, energy justice, food justice, environmental justice and urban inequality as described above.

Author contributions

Conceptualization-O.A., R.F., M.B. and G.M.; methodology-O.A., R.F., M.B. and G.M.; analysis-O.A. and R.F.; investigation-O.A., R.F., M.B. and G.M.; resources-R.F., O.A., M.B., F.T., T.N., G.M.; data curation- R.F., O.A., M.B., T.N., G.M.; writing—original draft preparation-O.A., R.F., F.T.; writing—review and editing- O.A., R.F., F.T., T.N., G.M. and M.B.; validation-O.A., R.F. and M.B., visualization-O.A., R.F., F.T.; supervision- R.F., M.B. and G.M.; project administration, R.F..; funding acquisition- R.F., M.B. and G.M. All authors have read and agreed to the

published version of the manuscript.

Funding

This research was funded by the Environmental Protection Agency and Health Service Executive of Ireland through the GBI-Health Research Project 2016-SE-DS-14.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: None

Data availability

I have shared the link to the data in the manuscript

Acknowledgement

The Authors wish to specially thank Professor Jan Rigby, formerly of the National Centre for Geocomputation, Maynooth University, Ireland (now retired); Dr. Conor Teljeur of the Chief scientist at Health Information and Quality Authority, Ireland; Dr. Aisling O'Connor of the Environmental Protection Agency of Ireland; Dr. Caitríona Carlin of the National University of Ireland Galway and NEAR Health programme; and Professor Mark Scott of University College Dublin and Eco-Health programme, Mary Morrissey, Declan McKeown and Fiona Donovan of the Health Service Executive of Ireland for their contributions and feedback to this research.

References

Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., 1998. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of International Conference on Management of Data of the Association for Computing Machinery's Special Interest Group on Management of Data, vol. 27. ACM SIGMOD) Conference 1998, New York, NY, United States, pp. 94–105. https://doi.org/10.1145/276305.276314, 2.

Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., 2005. Automatic subspace clustering of high dimensional data. Data Min. Knowl. Discov. 11, 5–33.

Alashwal, H., El Halaby, M., Crouse, J.J., Abdalla, A., Moustafa, A.A., 2019. The application of unsupervised clustering methods to alzheimer's disease. Front. Comput. Neurosci. 13, 31. https://doi.org/10.3389/fncom.2019.00031. PMID: 31178711; PMCID: PMC6543980.

Alhasoun, F., Aleissa, F., Alhazzani, M., Moyano, L.G., Pinhanez, C., Gonzalez, M.C., 2018. Age density patterns in patients medical conditions: a clustering approach. PLoS Comput. Biol. 14 (6), e1006115 https://doi.org/10.1371/journal.pcbi.1006115.

Alvioli, M., Marchesini, I., Reichenbach, P., Rossi, M., Ardizzone, F., Fiorucci, F., Guzzetti, F., 2016. Automatic delineation of geomorphological slope units with r. slopeunits v1.0 and their optimization for landslide susceptibility modeling. Geosci. Model Dev. (GMD) 9, 3975–3991. https://doi.org/10.5194/gmd-9-3975-2016.

Amoly, E., Dadvand, P., Forns, J., López-Vicente, M., Basagaña, X., Julvez, J., Alvarez-Pedrerol, M., Nieuwenhuijsen, M.J., Sunyer, J., 2014. Green and blue spaces and behavioral development in Barcelona schoolchildren: the BREATHE Project. Environ. Health Perspect. 122 (12), 1351–1358. https://doi.org/10.1289/ehp.1408215.

Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. In: Proceedings of International Conference on Management of Data of the Association for Computing Machinery's Special Interest Group on Management of Data. ACM SIGMOD) Conference 1999, Philadelphia PA, pp. 49–60. https://doi.org/10.1145/304182.304187.

Anselin, L., 1995. Local indicators of spatial association—LISA. Geogr. Anal. 27 (2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Anselin, L., 2005. Exploring Spatial Data with GeoDaTM: A Workbook. Spatial Analysis Laboratory, p. 138.

Arodudu, O.T., 2013. RS and GIS for City's environmental intelligence and precision planning. Aachener Geographische Arbeiten" (AGA) Series, Heft 50, 13–34.

Arodudu, O.T., Foley, R., Brennan, M., Mills, G., Bradley, M., Ningal, T., 2017. Green and Blue Infrastructure and Human Health. A Poster Presentation Presented on the September 15, 2017 at the. Yale Sustainability Leadership Forum, New Haven, Connecticut, USA.

Arodudu, O.T., Brennan, M., Mills, G., Ningal, T., Bradley, M., Foley, R., 2018. Health data within the Irish statistical system: identifying new sources and geographies. In: 50th Conference of Irish Geographers. Maynooth University, Ireland. May 8-10, 2018.

Arodudu, O.T., Foley, R., Brennan, M., Mills, G., Bradley, M., Ningal, T., 2019. Towards a more holistic framework for a health-led approach at the green and blue infrastructure and human health interface-Case study of Ireland. In: IALE (International Association of Landscape Ecology) World Congress. July 1-5, 2019, Milano, Italy.

Astell-Burt, T., Mitchell, R., Hartig, T., 2014. The association between green space and mental health varies across the lifecourse. A longitudinal study. J. Epidemiol. Community 68 (6), 578–583. https://doi.org/10.1136/jech-2013-203767.

Bell, S.L., Foley, R., Houghton, F., Maddrell, A., Williams, A.M., 2018. From therapeutic landscapes to healthy spaces, places and practices: a scoping review. Soc. Sci. Med. 196, 123–130, 1982.

Beyer, K.M.M., Kaltenbach, A., Szabo, A., Bogar, S., Nieto, F.J., Malecki, K.M., 2014. Exposure to neighborhood green space and mental health: evidence from the survey of the health of Wisconsin. Int. J. Environ. Res. Publ. Health 11, 3453–3472.

Bodenhofer, U., Kothmeier, A., Hochreiter, S., 2011. APCluster: an R package for affinity propagation clustering. Bioinformatics 27 (17), 2463–2464. https://doi.org/10.1093/bioinformatics/btr406.

Bullard, R.D., 2001. Environmental justice in the 21st century: race still matters. Phylon 49 (3–4), 151–171.

Bullard, R.D., 2003. Environmental justice for all. Crisis 110, 24.

Bulley, H.N.N., Arodudu, O.T., Obonyo, E.A., Polo-Akpisso, A., Ibrahim, E.S., Bamutaze, Y., 2023. Conservation planning in rapidly changing landscapes and disturbance regimes in the Global South. Int. J. Appl. Geospatial Res. (IJAGR) 14 (1), 1–23.

Calderón-Argelich, A., Benetti, S., Anguelovski, I., Connolly, J.J.T., Langemeyer, J., Baró, F., 2021. Tracing and building up environmental justice considerations in the urban ecosystem service literature. A systematic review, Landscape and Urban Planning 214, 104130.

Calogiuri, G., Chroni, S., 2014. The impact of the natural environment on the promotion of active living: an integrative systematic review. BMC Public 14, 873.

Campello, R.J.G.B., Kröger, P., Sander, J., Zimek, A., 2020. Density-based clustering. WIREs Data Mining Knowledge Discovery 10, e1343. https://doi.org/10.1002/widm.1343.

Celebi, M.E., Aslandogan, Y.A., Bergstresser, P.R., 2005. Mining biomedical images with density-based clustering. In: International Conference on Information Technology: Coding and Computing, vol. II. ITCC'05), pp. 163–168.

Charest, L., Plante, J.-F., 2014. Using balanced iterative reducing and clustering hierarchies to compute approximate rank statistics on massive datasets. J. Stat. Comput. Simulat. 84 (10), 2214–2232. https://doi.org/10.1080/00949655.2013.787534.

Chiu, T., Fang, D., Chen, J., Wang, Y., Jeris, C., 2001. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 263–268. https://doi.org/10.1145/502512.502549.

Cui, J., Wang, Y., Wang, K., 2022. Key technology of the medical image wise mining method based on the meanshift algorithm. Emergency Medicine International, 6711043. https://doi.org/10.1155/2022/6711043. PMID: 35757271; PMCID: PMC9217612.

Delil, S., Çelik, R.N., San, S., Dundar, M., 2017. Clustering patient mobility patterns to assess effectiveness of health-service delivery. BMC Health Serv. Res. 17 (1), 458. https://doi.org/10.1186/s12913-017-2381-2. PMID: 28676090; PMCID: PMC5497378.

DM365, 2020. Grid-based clustering - STING, WaveCluster & CLIQUE. Data Mining 365. Retrieved from. https://www.datamining365.com/2020/04/grid-based-clustering.html. (Accessed 1 December 2022).

Dunning, C., 2020. An Overview of Hierarchical Cluster Analysis (HCA). Data Science Student Society, UC San Diego, California, United States. Retrieved from. https://medium.com/ds3ucsd/an-overview-of-hierarchical-cluster-analysis-hca-84f37f99bc7c. (Accessed 1 December 2022).

ESRI, 2020a. How Do You Represent NULL Values when Converting from a Geodatabase, or Shapefile, to a Coverage? Retrieved from. https://support.esri.com/en-us/knowledge-base/faq-how-do-you-represent-null-values-when-converting-fr-000004419 on August 2, 2023.

ESRI, 2020b. Data Represented by Null Values Are Not Symbolized on the Map in ArcGIS Online. Retrieved from. https://support.esri.com/en-us/knowledge-base/problem-data-represented-by-null-values-are-not-symboli-000017879 on August 2, 2023.

Fichtenberger, H., Gillé, M., Schmidt, M., Schwiegelshohn, C., Sohler, C., 2013. BICO: BIRCH meets coresets for k-means clustering. In: Bodlaender, H.L., Italiano, G.F. (Eds.), ESA 2013. LNCS, 8125. Springer, Heidelberg, pp. 481–492. https://doi.org/10.1007/978-3-642-40450-4_41.

Foley, R., 2015. Swimming in Ireland: immersions in therapeutic blue space. Health Place 35, 218–225.

Foley, R., Kistemann, T., 2015. Blue space geographies: enabling health in place. Health Place 35, 157–165.

Foley, R., Brennan, M., Arodudu, O., Mills, G., Ningal, T., Bradley, M., 2018a. Green and Blue Spaces and Health: A Health-Led Approach, vol. 264. EPA Synthesis Report, Research, p. 73.

Foley, R., Arodudu, O., Brennan, M., Mills, G., Bradley, M., Ningal, T., 2018b. GBI Health Policy Brief I: Finding the Right Geography for Health Data for Ireland, p. 6.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. Science 315 (5814), 972–976. https://doi.org/10.1126/science.1136800. Epub 2007 Jan 11. PMID: 17218491.

Frumkin, H., 2003. Healthy places: exploring the evidence. Am. J. Publ. Health 93 (9), 1451–1456.

Fuller, R.A., Irvine, K.N., Devine-Wright, P., Warren, P.H., Gaston, K.J., 2007. Psychological benefits of greenspace increase with biodiversity. Biol. Lett. 3, 390–394.
Gardner-Frolick, R., Boyd, D., Giang, A., 2022. Selecting data analytic and modeling methods to support air pollution and environmental justice investigations: a critical review and guidance framework. Environ. Sci. Technol. 56 (5), 2843–2860. https://doi.org/10.1021/acs.est.1c01739.
Gascon, M., Triguero-Mas, M., Martínez, D., Dadvand, P., Forns, J., Plasència, A., Nieuwenhuijsen, M.J., 2015. Mental health benefits of long-term exposure to residential green and blue spaces: a systematic review. Int. J. Environ. Res. Publ. Health 12, 4354–4379.
Getis, A., 2010. The analysis of spatial association by use of distance statistics. Geogr. Anal. 24 (3), 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.
Gonzales, A., 2022. Find Out Your Community's Environmental Justice Score. Retrieved from. https://salud-america.org/find-out-your-communitys-environmental-justice-score/. (Accessed 19 April 2023).
Grieve, J., 2011. A regional analysis of contraction rate in written Standard American English. Int. J. Corpus Linguist. 16 (4), 514–546. https://doi.org/10.1075/ijcl.16.4.04gri.
Groenewegen, P.P., van den Berg, A.E., de Vries, S., Verheij, R.A., 2006. Vitamin G: effects of green space on health, well-being, and social safety. BMC Publ. Health 6, 149.
Gupta, A., 2021. Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH) Algorithm in Machine Learning. Geek Culture. Retrieved from. https://medium.com/geekculture/balanced-iterative-reducing-and-clustering-using-hierarchies-birch-1428bb06bb38. (Accessed 1 December 2022).
Han, J., Pei, J., Tong, H., 2022. Data Mining: Concepts and Techniques. Morgan Kaufmann.
Hartig, T., Evans, G.W., Jamner, L.D., Davis, D.S., Gärling, T., 2003. Tracking restoration in natural and urban field settings. J. Environ. Psychol. 23, 109–123.
Helbich, M., Leitner, M., Kapusta, N.D., 2012. Geospatial examination of lithium in drinking water and suicide mortality. Int. J. Health Geogr. 11 (1), 19. https://doi.org/10.1186/1476-072X-11-19. PMC 3441892. PMID 22695110.
Heo, S., Bell, M.L., 2023. Investigation on urban greenspace in relation to sociodemographic factors and health inequity based on different greenspace metrics in 3 US urban communities. J. Expo. Sci. Environ. Epidemiol. 33 (2), 218–228.
Heo, S., Lim, C.C., Bell, M.L., 2020. Relationships between local green space and human mobility patterns during COVID-19 for Maryland and California, USA. Sustainability 12, 9401.
Hicks, S.C., Liu, R., Ni, Y., Purdom, E., Risso, D., 2021. mbkmeans: fast clustering for single cell data using mini-batch k-means. PLoS Comput. Biol. 17 (1), e1008625 https://doi.org/10.1371/journal.pcbi.1008625.
Ibrahim, E.S., Ahmed, B., Arodudu, O.T., Abubakar, J.B., Dang, B.A., Mahmoud, M.I., Shaba, H.A., Shamaki, S.B., 2022. Desertification in the sahel region: a product of climate change or human activities? A case of desert encroachment monitoring in north-eastern Nigeria using remote sensing techniques. Geographies 2 (2), 204–226. https://doi.org/10.3390/geographies2020015.
Jianjun, L., Jianquan, K., 2018. Recognition of genetically modified product based on affinity propagation clustering and terahertz spectroscopy. Spectrochim. Acta Mol. Biomol. Spectrosc. 194, 14–20. https://doi.org/10.1016/j.saa.2017.12.074.
Joshi, S., 2022. What Is Clustering in Machine Learning: Types and Methods. Retrieved from. https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms. (Accessed 1 December 2022).
Kassambara, A., 2019. Agglomerative hierarchical clustering, hierarchical clustering in R: the essentials, datanovia. Retrieved from. https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/. (Accessed 1 December 2022).
Kindermann, G., Domegan, C., Britton, E., Carlin, C., Mashinchi, M.I., Ojo, A., 2021. Understanding the dynamics of green and blue spaces for health and wellbeing outcomes in Ireland: a systemic stakeholder perspective. Sustainability 13, 9553.
Kriegel, H.-P., Kröger, P., Sander, J., Zimek, A., 2011. Density-based clustering. WIREs Data Mining Knowledge Discovery 231–240. https://doi.org/10.1002/widm.30.
Lachowycz, K., Jones, A.P., 2013. Towards a better understanding of the relationship between greenspace and health: development of a theoretical framework. Landsc. Urban Plann. 118, 62–69.
Lang, A., Schubert, E., 2020. BETULA: Numerically Stable CF-Trees for BIRCH Clustering, Similarity Search and Applications. Retrieved from, pp. 281–296. https://doi.org/10.1007/978-3-030-60936-8_22. https://link.springer.com/chapter/10.1007/978-3-030-60936-8_22. (Accessed 1 August 2023).
Lang, A., Schubert, E., 2022. BETULA: fast clustering of large data with improved BIRCH CF-Trees. Inf. Syst. 108, 101918 https://doi.org/10.1016/j.is.2021.
Li, H., Calder, C.A., Cressie, N., 2007. Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. Geogr. Anal. 39 (4), 357–375. https://doi.org/10.1111/j.1538-4632.2007.00708.x.
Loftus, T.J., Shickel, B., Balch, J.A., Tighe, P.J., Abbott, K.L., Fazzone, B., Anderson, E. M., Rozowsky, J., Ozrazgat-Baslanti, T., Ren, Y., Berceli, S.A., Hogan, W.R., Efron, P. A., Moorman, J.R., Rashidi, P., Upchurch Jr., G.R., Bihorac, A., 2022. Phenotype clustering in health care: a narrative review for clinicians. Frontiers in Artificial Intelligence 5, 842306. https://doi.org/10.3389/frai.2022.842306. PMID: 36034597; PMCID: PMC9411746.
Maas, J., Verheij, R.A., Groenewegen, P.P., De Vries, S., Spreeuwenberg, P., 2006. Green space, urbanity, and health: how strong is the relation? J. Epidemiol. Community 60 (7), 587–592.
Martin, C., 2021. The Landscape of Data Capacity in US Environmental Justice Organizations. Urban, Institute, The Brookings Institution and Harvard Joint Center for Housing Studies, p. 45.
Martuzzi, M., Mitis, F., Forastiere, F., 2010. Inequalities, inequities, environmental justice in waste management and health. Eur. J. Publ. Health 20 (1), 21–26. https://doi.org/10.1093/eurpub/ckp216.
McFarlane, R.A., Sleigh, A.C., McMichael, A.J., 2013. Land-use change and emerging infectious disease on an island continent. Int. J. Environ. Res. Publ. Health 10 (7), 2699–2719. Find Out Your Community's Environmental Justice Score.
Menton, M., Larrea, C., Latorre, S., Martinez-Alier, J., Peck, M., Temper, L., Walter, M., 2020. Environmental justice and the SDGs: from synergies to gaps and contradictions. Sustain. Sci. 15, 1621–1636.
Mitchell, R., Astell-Burt, T., Richardson, E.A., 2011. A comparison of green space indicators for epidemiological research. J. Epidemiol. Community 65, 853–858.
Mitchell, R.J., Richardson, E.A., Shortt, N.K., Pearce, J.R., 2015. Neighborhood environments and socioeconomic inequalities in mental well-being. Am. J. Prev. Med. 49 (1), 80–84. https://doi.org/10.1016/j.amepre.2015.01.017.
Moreira, A., Santos, M.Y., 2005. 1 Density-Based Clustering Algorithms – DBSCAN and SNN.
Mueller, N., Rojas-Rueda, D., Basagaña, X., Cirach, M., Cole-Hunter, T., Dadvand, P., Donaire-Gonzalez, D., Foraster, M., Gascon, M., Martinez, D., Tonne, C., Triguero-Mas, M., Valentín, A., Nieuwenhuijsen, M., 2016. Urban and transport planning related exposures and mortality: a health impact assessment for cities. Environ. Health Perspect. 125 (1), 89–96. https://doi.org/10.1289/EHP22.
Neill, D.B., Moore, A., 2004. Fast grid-based scan statistic for detection of significant spatial disease cluster. MMWR Supplement 53 (Suppl. l), 255.
NIST, 2012. Introduction to Time Series Analysis: Definitions, Applications and Techniques, Process or Product Monitoring and Control. Retrieved from. https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc41.htm. (Accessed 1 December 2022).
Ogbodo, J.A., Egbuche, C.T., Arodudu, O.T., 2014. Detecting logging roads and clearcuts with TerraSAR and RADARSAT data. In: Lac, S., McHenry, M.P. (Eds.), Climate Change and Forest Ecosystems. Nova Science Publishers, Hauppauge, NY, pp. 71–88.
Pearson, A.L., Bentham, G., Day, P., Kingham, S., 2014. Associations between neighbourhood environmental characteristics and obesity and related behaviours among adult New Zealanders. BMC Publ. Health 14, 553–566.
Pellicer-Valero, O.J., Fernández-de-las-Peñas, C., Martín-Guerrero, J.D., Navarro-Pardo, E., Cigarán-Méndez, M.I., Florencio, L.L., 2020. Patient profiling based on spectral clustering for an enhanced classification of patients with tension-type headache. Appl. Sci. 10 (24), 9109. https://doi.org/10.3390/app10249109.
Pu, Q., Yoo, E.-H., Rothstein, D.H., Cairo, S., Malemo, L., 2020. Improving the spatial accessibility of healthcare in north kivu, democratic republic of Congo. Appl. Geogr. 121, 102262 https://doi.org/10.1016/j.apgeog.2020.102262. ISSN 0143-6228.
Rabel, M., Laxy, M., Thorand, B., Peters, A., Schwettmann, L., Mess, F., 2019. Clustering of health-related behavior patterns and demographics, results from the population-based KORA S4/F4 cohort study. Front. Public Health 6, 387. https://doi.org/10.3389/fpubh.2018.00387. PMID: 30723712; PMCID: PMC6350271.
Ramadhani, F., Zarlis, M., Suwilo, S., 2019. Improve BIRCH algorithm for big data clustering. In: IOP Conference Series: Materials Science and Engineering, vol. 725, 012090.
Razzak, M.I., Imran, M., Xu, G., 2020. Big data analytics for preventive medicine. Neural Comput. Appl. 32, 4417–4451. https://doi.org/10.1007/s00521-019-04095-y, 2020.
Rigby, J.E., Boyle, M.G., Brunsdon, C., Charlton, M., Dorling, D., French, W., Pringle, D., 2017. Towards a geography of health inequalities in Ireland. Ir. Geogr. 50 (1), 1–27.
Roman, R.A., 2021. Application of spectral clustering for the detection of high priority areas of attention for COVID-19 in Mexico. In: Marmolejo-Saucedo, J.A., Vasant, P., Litvinchev, I., Rodriguez-Aguilar, R., Martinez-Rios, F. (Eds.), Computer Science and Health Engineering in Health Services. COMPSE 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 359. Springer, Cham. https://doi.org/10.1007/978-3-030-69839-3_9.
Sanders, T., Feng, X., Fahey, P.P., Lonsdale, C., Astell-Burt, T., 2015. Greener neighbourhoods, slimmer children? Evidence from 4423 participants aged 6 to 13 years in the Longitudinal Study of Australian children. Int. J. Obes. 39 (8), 1224–1229. https://doi.org/10.1038/ijo.2015.69.
Santhanam, T., Padmavathi, M.S., 2014. Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Comput. Sci. 47 (C), 76–83. https://doi.org/10.1016/j.procs.2015.03.185.
Schultz, T., Kindlmann, G.L., 2013. Open-box spectral clustering: applications to medical image analysis. IEEE Trans. Visual. Comput. Graph. 19 (12), 2100–2108. https://doi.org/10.1109/TVCG.2013.181. PMID: 24051776.
Sculley, D., 2010. Web-scale k-means clustering. WWW 10. In: Proceedings of the 19th International Conference on World Wide Web. Association for Computing Machinery, New York, NY, USA, pp. 1177–1178. https://doi.org/10.1145/1772690.1772862.
Shuai, Y., 2022. A full-sample clustering model considering whole process optimization of data. Big Data Research 28, 100301.
Smith, G., Cirach, M., Swart, W., Dėdelė, A., Gidlow, C., van Kempen, E., Kruize, H., Gražulevičienė, R., Nieuwenhuijsen, M.J., 2017. Characterisation of the natural environment: quantitative indicators across Europe. Int. J. Health Geogr. 16 (1), 16–31.
Song, H., Lewis, N.A., Ballew, M.T., Bravo, M., Davydova, J., Gao, H.O., Garcia, R.J., Hiltner, S., Naiman, S.M., Pearson, A.R., Romero-Canyas, R., Schuldt, J.P., 2020. What counts as an issue? Differences in issue conceptualization by race, ethnicity, and socioeconomic status. J. Environ. Psychol. 68, 101404.
Taghikhah, F., Voinov, A., Shukla, N., Filatova, T., 2020. Exploring consumer behavior and policy options in organic food adoption: insights from the Australian wine sector. Environ. Sci. Pol. 109, 116–124.

Taghikhah, F., Voinov, A., Shukla, N., Filatova, T., 2021. Shifts in consumer behavior towards organic products: theory-driven data analytics. J. Retailing Consum. Serv. 61, 102516.

Taghikhah, F., Erfani, E., Bakhshayeshi, I., Tayari, S., Karatopouzis, A., Hanna, B., 2022a. Artificial intelligence and sustainability: solutions to social and environmental challenges. In: Artificial Intelligence and Data Science in Environmental Sensing. Elsevier, pp. 93–108.

Taghikhah, F., Voinov, A., Filatova, T., Polhill, J.G., 2022b. Machine-assisted agent-based modeling: opening the black box. Journal of Computational Science 64, 101854.

Thanniru, S., 2022. Introduction to Spectral Clustering, Data Science and Business Analytics. Retrieved from. https://www.mygreatlearning.com/blog/introduction-to-spectral-clustering/. (Accessed 1 December 2022).

Tripathi, A., 2022. Mean-Shift Clustering. Retrieved from. https://www.geeksforgeeks.org/ml-mean-shift-clustering/. (Accessed 1 December 2022).

Völker, S., Kistemann, T., 2011. The impact of blue space on human health and wellbeing–Salutogenetic health effects of inland surface waters: a review. Int. J. Hyg Environ. Health 214, 449–460.

Völker, S., Kistemann, T., 2014. Developing the urban blue: comparative health responses to blue and green urban open spaces in Germany. Health Place 35, 196–205.

Wang, W., Yang, J., Muntz, R.R., 1997. STING: a statistical information grid approach to spatial data mining, VLDB '97. In: Proceedings of the 23rd International Conference on Very Large Data Bases, pp. 186–195.

Waqas, S.M., Hussain, K., Mostafa, S.A., Nawi, N., Khan, S., 2022. Fuzzy density-based clustering for medical diagnosis. In: Ghazali, R., Mohd Nawi, N., Deris, M.M., Abawajy, J.H., Arbaiy, N. (Eds.), Recent Advances in Soft Computing and Data Mining. SCDM 2022, Lecture Notes in Networks and Systems, vol. 457. Springer, Cham. https://doi.org/10.1007/978-3-031-00828-3_26.

Wardani, R.S., Sayono, P., Aditya Paramananda, A., 2019. Clustering tuberculosis in children using K-Means based on geographic information system. AIP Conf. Proc. 2114, 060012 https://doi.org/10.1063/1.5112483.

Wheeler, B.W., White, M., Stahl-Timmins, W., Depledge, M.H., 2012. Does living by the coast improve health and wellbeing? Health Place 18 (5), 1198–1201.

Wheeler, B.W., Lovell, R., Higgins, S.L., White, M.P., Alcock, I., Osborne, N.J., Husk, K., Sabel, C.E., Depledge, M.H., 2015. Beyond greenspace: an ecological study of population general health and indicators of natural environment type and quality. Int. J. Health Geogr. 14, 17. https://doi.org/10.1186/s12942-015-0009-5, 2015.

White, M., Smith, A., Humphryes, K., Pahl, S., Snelling, D., Depledge, M., 2010. Blue space: the importance of water for preference, affect, and restorativeness ratings of natural and built scenes. J. Environ. Psychol. 30 (4), 482–493.

Wilker, E.H., Wu, C.D., McNeely, E., Mostofsky, E., Spengler, J., Wellenius, G.A., Mittleman, M.A., 2014. Green space and mortality following ischemic stroke. Environ. Res. 133, 42–48. https://doi.org/10.1016/j.envres.2014.05.005.

Wing, S., 2005. Environmental justice, science, and public health. Essays on the Future of Environmental Health Research 54–63.

Xiao, B., Wang, Z., Liu, Q., Liu, X., 2018. SMK-Means: an Improved Mini Batch K-Means Algorithm Based on Mapreduce with Big Data. Tech Science Press, pp. 1–5, 1.

Zhang, T., Ramakrishnan, R., Livny, M., 1995. BIRCH: an Efficient Data Clustering Method for Very Large Databases, Technical Report. Computer Sciences Dept., Univ. of Wisconsin-Madison.

Zhang, T., Ramakrishnan, R., Livny, M., 1996. BIRCH: an efficient data clustering method for very large databases. ACM Sigmod Record, ACM 25, 103–114.

Zhang, T., Ramakrishnan, R., Livny, M., 1997. BIRCH: a new data clustering algorithm and its applications. WIREs Data Mining Knowledge Discovery 1 (2), 141–182. https://doi.org/10.1023/A:1009783824328.

Zhang, T., Ramakrishnan, R., Livny, M., 1999. Fast density estimation using CF-kernel for very large databases. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 312–316. https://doi.org/10.1145/312129.312266.

Zhang, J., Liu, Q., Chen, H., Yuan, Z., Huang, J., Deng, L., Lu, F., Zhang, J., Wang, Y., Wang, M., Chen, L., 2015. Combining self-organizing mapping and supervised affinity propagation clustering approach to investigate functional brain networks involved in motor imagery and execution with fMRI measurements. Front. Hum. Neurosci. 9, 400. https://doi.org/10.3389/fnhum.2015.00400.

Zhang, Z., Murtagh, F., Van Poucke, S., Lin, S., Lan, P., 2017. Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting visualization with R. Ann. Transl. Med. 5 (4), 75. https://doi.org/10.21037/atm.2017.02.05. PMID: 28275620; PMCID: PMC5337204.