The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume L-4/W1-2026
Free and Open Source Software for Geospatial (FOSS4G) 2026 – Academic Track "Bridging Geospatial Technology and Humanity",
30 August–5 September 2026, Hiroshima, Japan

Data4Land - Reproducible Open-Source Tool for Enrichment of Land Use / Land Cover Rasters and Connectivity Maps

Vitalii Kriukov [1], Lucy Bastin [2], Riyad Rahman [3]
[1] School of Computer Science and Digital Technologies, Aston University, Birmingham, UK -vit.kryukov@gmail.com
[2] Aston Centre for Artificial Intelligence Research and Application, Aston University, Birmingham, UK - l.bastin@aston.ac.uk
[3] School of Computer Science and Digital Technologies, Aston University, Birmingham, UK - riyad23008@gmail.com

Keywords: Land Use, Habitat Connectivity, Spatial Planning, OpenStreetMap, Python, Command-Line Tool

Abstract

Land use/ land cover (LULC) datasets derived from remote sensing are widely used in geospatial applications for environmental management, natural capital assessment, and spatial planning. However, their spatial resolution, thematic consistency, and accuracy are often insufficient for complex analyses including landscape connectivity, fragmentation metrics and spatial inequality assessments. This paper presents Data4Land, an open-source Python-based workflow that systematically enriches land use/land cover datasets with auxiliary vector data, such as OpenStreetMap and the World Database on Protected Areas. The tool is demonstrated for two contrasting case study areas, using multi-temporal LULC time series: Catalonia in Spain (2012-2022) and Northern England (2020-2023). Integrating road, railway, watercourse and protected area features substantially changed habitat connectivity indices at multiple scales. Enrichment accuracy was validated against Ordnance Survey vector data using a confusion matrix approach. Data4Land is distributed as a modular Python package with Docker containerisation and is freely available on GitHub.

1. Introduction

1.1. Land Use and Land Cover in Geospatial Studies

Land use and land cover (LULC) datasets derived from satellite remote sensing are foundational inputs for a broad range of environmental applications, including spatial planning, deforestation and habitat degradation monitoring, ecosystem service assessment, agricultural tracking, and urban footprint analysis (Wang et al, 2023). Despite continuous advances in Earth observation technology and the proliferation of open-access products such as ESA WorldCover and Corine Land Cover, these datasets remain imperfect: they can be spatially incomplete, thematically inconsistent, or insufficiently detailed for the specific ecological or planning context in which they are applied. One recurring and ecologically consequential limitation is the systematic underrepresentation of narrow linear infrastructure, particularly roads and railways, in widely-used medium-resolution products such as Landsat or Sentinel-2. The detection of these features is limited not only by pixel size, but also by spectral confusion arising from shadows cast by adjacent vegetation canopy (Immitzer et al, 2016). Such infrastructure dissects natural and semi-natural habitats, acting as barriers to wildlife movement and fundamentally altering the functional connectivity of landscapes (Cork et al, 2024).

Accurate and spatially consistent LULC datasets are a prerequisite for tools used in ecological studies, particularly those concerned with habitat connectivity and species dispersal modelling (Pechanec et al, 2025). A central derived metric in such analyses is landscape resistance, also termed impedance or friction, which characterises the difficulty experienced by individuals of a given species when migrating between different land cover types, such as urban hard surfaces, open water, or grassland (Beier et al, 2008; Keely et al, 2021). Resistance surfaces are typically parameterised from LULC data, meaning that omissions or errors in the underlying map propagate directly into connectivity model outputs (Zeller et al, 2012; Prima et al, 2024). When linear barriers such as roads are absent from the LULC input, the modelled landscape falsely appears more permeable than it is in reality. This bears potentially significant consequences for conservation planning decisions (Brennan et al, 2025). Beyond connectivity analysis, enriched LULC datasets have demonstrated utility in modelling future land conversion trajectories and in quantifying land surface temperature dynamics and urban heat island effects (Mahamune and Gavankar, 2025).

1.2. OpenStreetMap Data for Land Use and Land Cover Analysis

OpenStreetMap (OSM) represents a compelling complementary data source for addressing many of these limitations. As a global, continuously updated, open-access volunteered geographic information (VGI) platform, OSM describes an exceptionally wide range of spatial features. This includes road networks, railways, water bodies, protected areas, and land use polygons at a thematic and geometric detail that medium-resolution satellite products cannot replicate (Haklay, Weber, 2008). Studies have demonstrated that OSM-derived land cover products can outperform established satellite-based datasets such as Corine Land Cover in thematic detail and classification accuracy, especially in heterogeneous landscapes (Shultz et al, 2017). The historical contribution logs of OSM, accessible via APIs such as Ohsome, enable the reconstruction of feature geometries at past timestamps, opening the possibility of producing temporally consistent enriched LULC time series, which has received comparatively little attention in the literature.

Despite these strengths, OSM data quality varies geographically, with urban and well-mapped regions generally exhibiting higher completeness and positional accuracy than rural or data-sparse areas (Viana et al, 2019), a consideration that users must account for when integrating OSM features into analytical workflows. Some prior studies have applied preprocessing strategies similar in spirit to LULC enrichment. These include work on improving habitat connectivity inputs (Sahraoui et al, 2017), as well as enrichment for non-ecological purposes (Patriarca et al, 2019). However, existing tools have notable shortcomings. The OSM LandUse interactive web-GIS tool (Shultz et al, 2017), for example, employs fixed data extraction parameters and a rigid layer-writing structure, and does not distinguish ecologically important linear barriers such as road networks. More broadly, the flexibility to adapt extraction logic, spatial buffering, and feature prioritisation to species- or taxa-specific requirements has been lacking. Integration of temporal OSM snapshots to generate

https://doi.org/10.5194/isprs-archives-L-4-W1-2026-135-2026

historically consistent enriched LULC maps has also not been particularly explored. This is an important capability for studies that analyse landscape change over time or seek to retrospectively model species responses to infrastructure development.

1.3. Data4Land for LULC Enrichment

In this paper, we present Data4Land, a semi-automatic, open-source tool that addresses these limitations by enabling the enrichment of raster LULC datasets with external vector sources of higher spatial accuracy and thematic consistency, primarily OSM. Originally developed to reintroduce narrow linear features. such as roads and railways, that are systematically absent from medium-resolution LULC products, Data4Land has been designed with sufficient flexibility to accommodate a wide range of feature types and taxa-specific use cases. Users interact with the tool through an interactive Python-based command line tool which extracts required information from CSV attribute tables and YAML configuration files. The tool interfaces with both the Ohsome and Overpass APIs for OSM data retrieval and supports integration of the UNEP-WCMC World Database on Protected Area (WDPA)s (2025) for users with Protected Planet API access. Crucially, Data4Land supports multi-temporal enrichment workflows, allowing LULC time series to be enriched with OSM features that existed at each corresponding point in time, enabling the construction of temporally consistent historical datasets. The Data4Land tool also allows users to derive landscape resistance surfaces directly from enriched outputs by configuring ecologically meaningful parameters, including the spatial decay of stressor effects, tailored to specific species or movement processes.

2. Methodology and Results

Data4Land v2.0 is implemented as a modular Python command-line tool comprising four processing blocks, each targeting a distinct stage of the LULC enrichment pipeline. The tool has been developed and validated against two contrasting case study areas: the Catalonia region of Spain and a part of Northern England.

2.1. Data4Land Architecture

Data4Land v2.0 includes the following modular blocks (Figure 1):

1. Ingestion of protected areas (*process-wdpa*)

2. Historical data ingestion (*process-osm*)

3. LULC enrichment (*enrich-lulc*)

4. Landscape resistance and stressors (*recalc-impedance*)

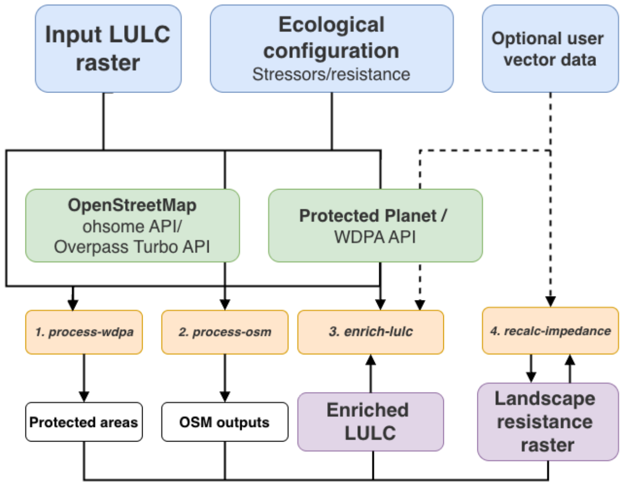


Figure 1. Conceptual modular framework of Data4Land tool.

Version 2.0 of Data4Land introduced a breaking change in architecture and functionality, replacing Notebook structure with a command-line mode built on modular Python classes (Kriukov et al, 2025). Aside from this, a few other significant changes have been implemented:

- Ohsome API to fetch OpenStreetMap data and improve performance compared to Overpass Turbo API;
- introduced option to run commands for series of input datasets covering multiple years;
- increased scalability with two sample datasets of different spatial extent and support of multiple case studies and sub-case studies.

Data4Land requires a Docker container to be built based on Python 3.12+, which includes common preinstalled geospatial libraries, such as gdal, ogr, osr, (geo)pandas, rasterio, fiona, numpy, and npm package manager.

It is important to note that the same LULC dataset can be used for multiple analyses within a single case study area. For example, a user may have LULC data for Catalonia covering the period 2012-2022 but wish to run separate analyses using different landscape resistance parameters. In such cases, a dedicated set of landscape resistance datasets must be provided for each sub-case study and specified in the configuration file.

For example, users may define sub-case studies based on LULC classes (*subcase_study:'forest'* or *subcase_study:'meadows'* to generate outputs for different land cover types. Sub-case studies can also represent other analysis targets, such as species, for example, *subcase_study: 'turtle'.*

More detailed documentation on usage and command examples is provided in the GitHub repository [1].

2.2. Case Studies

The first case study area covers the Catalonia region in Spain, (an area of 32,110 km^2), across a 2012-2022 time series. The original land use / land cover with a spatial resolution of 30 m dataset provides a detailed classification with 7 types. The following additional features were ingested from OSM: roads in a number of size categories, railways, water bodies, waterways and vineyards. Protected areas were also ingested from WDPA (further discussed in section 2.4). The enriched land use / land

[1] https://github.com/AD4GD/pilot-2-preprocessing

https://doi.org/10.5194/isprs-archives-L-4-W1-2026-135-2026 |

cover datasets were then used as an input data for habitat connectivity calculations using the Graphab software (Foltête J.-C, 2021).

The input LULC data used is the Land Use / Land Cover Map of Catalonia (MUCSC) created by CREAF, UAB and the Generalitat de Catalunya (Gonzales-Guerrero and Pons, 2020). This collection is produced through automatic classification of Landsat and Sentinel images for the Generalitat de Catalunya. To avoid ‘edge effects’ and clipping of output data along image borders in follow-up calculations, MUCSC was combined with three ancillary datasets covering adjacent regions (Bastin et al., 2024).

The second case study area covers a part of Northern England covering 2,120 km^2, including parts of Manchester and Sheffield cities, as well as the majority of the Peak District National Park. This case study area covers a wide range of natural, semi-natural and human-centered landscapes (2020-2023 time series). 21 LULC classes were available from the UK Centre for Ecology and Hydrology Land Cover annual maps (Morton et al, 2024) which have 10 m spatial resolution (Figure 2). Again, OSM features as above (excluding vineyards) and WDPA protected areas were retrieved to enrich the LULC map.

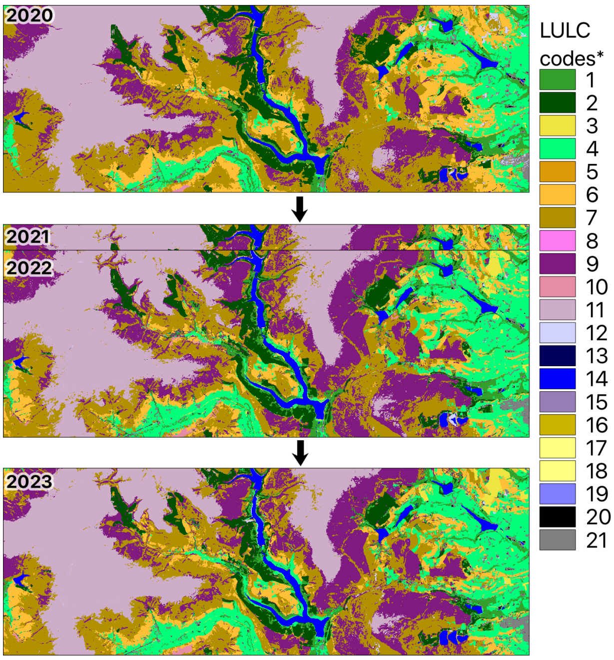


Figure 2. Land use/land cover timeseries for a part of Northern England, 2020-2023 (*type codes are given according to Morton et al, 2024).

The two case studies’ LULC data sources were obtained through different processing workflows, and the OSM features have been constructed from a range of authoritative, academic, or citizen science sources which significantly vary between the countries.

Both areas of interest are characterised by a wide range of geographic conditions and habitats. Catalonia includes Mediterranean, montane, and alpine ecosystems characterised by high climatic and elevational heterogeneity (Torras and Saura, 2009), while the Peak District is dominated by temperate upland habitats such as heather moorland, blanket bog, and species-rich grasslands, gradually falling to varying grasslands mixed with suburban settlements (Peak District, 2021).

2.3. OpenStreetMap Ingestion

In the version 1.1, the *process-osm* command uses Overpass Turbo API to retrieve historical OpenStreetMap (OSM) data through the Overpass Query Language (Overpass QL). The API enables extraction of selected OSM elements, including nodes, ways, and relations, that match predefined filtering criteria.

The retrieved data are returned in JSON format and include all spatial features satisfying the specified OSM tag filters within the bounding box of the input raster dataset. Data is extracted for a user-defined reference year provided in the configuration file. Historical OSM data are available from September 2012 onwards; however, data consistency generally improves after 2013, making earlier records less suitable for analysis.

To support LULC enrichment, four query groups were developed to extract roads, railways, waterways, and water bodies. The filtering rules were derived from previous studies, OSM tagging guidelines, and extensive testing of feature relevance for LULC refinement.

Users should be aware of several limitations associated with historical OSM data. These are primarily related to missing information, tagging inconsistencies, and temporal changes in mapping practices rather than geometric or topological errors. Defining suitable tag filters can be particularly challenging for historical analyses because most OSM documentation focuses on current tagging conventions. In addition, the use of specific tags has evolved over time. For example, the classification of aerial water features has undergone substantial changes, and tagging practices may differ considerably between regions (Kriukov et al, 2025).

To accommodate these variations, Data4Land allows users to define custom combinations of OSM keys and values for data extraction. This flexibility enables query definitions to be adapted to the geographical context, study objectives, and characteristics of the available historical OSM data.

In the version 2.0, two APIs are available to use in the command-line mode to extract OSM features – Ohsome API and Overpass Turbo API. Both APIs have been tested for both case study areas, and they fetch consistent sets of features (Table 1). Although Overpass Turbo API provides more comprehensive documentation, Ohsome API has been proved to be significantly faster for scales such as Catalonia or Northern England (Table 2). In general, it is recommended to use Ohsome API, as it is more stable and provides quicker fetching, especially for multiple timestamps and large spatial extents.

Area and year	1 (Catalonia)	2 (Northern England)
roads	89,690	11,750
railways	6,819	1,647
water lines	3,112	611
water bodies	21,541	2,165
vineyards	8,408	n/a

Table 1. Number of features fetched for both case studies (2022)

Area and year	Ohsome API, s	Overpass Turbo API, s
UK (2020)	16.10	335.05
UK (2021)	18.25	396.88
UK (2022)	18.41	380.30
UK (2023)	19.25	494.45
Catalonia (2012)	38.02	385.08
Catalonia (2017)	41.54	561.59
Catalonia (2022)	57.29	630.86

Table 2. Performance benchmarking of Ohsome API vs Overpass Turbo API (roads, railways, water bodies, and water lines)

Although Ohsome API provides better performance, the Overpass Turbo option has been retained to permit flexibility. In version 2.0 of Data4Land, some queries have been enhanced, for example, accounting for tunnels and bridges which provide non-planar links in the landscape and thus are crucial for environmental research and habitat connectivity assessments (Figure 3).

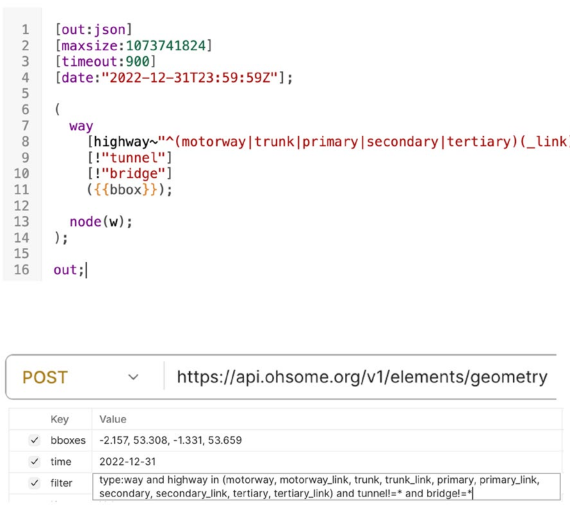


Figure 3. Query example for road features in OSM, using Overpass Turbo API (above) and Ohsome API (below).

2.4. World Database on Protected Areas Ingestion

Protected areas (PAs) play a key role in biodiversity conservation by safeguarding threatened species and improving habitat connectivity. For this reason, information on the spatial extent of protected areas is commonly incorporated into conservation analyses, as they may act as refugia or decrease the effective resistance of the landscape to species' movements.

Data4Land retrieves PA boundaries that overlap the bounding box of the input raster dataset through the Protected Planet API. Access to this service requires authentication using an access token which should be defined by the user in the configuration YAML file.

The retrieved PA data, provided in JSON format, undergo a harmonisation process that includes merging paginated API responses, filtering protected areas based on the acquisition date of the input LULC dataset and the year of PA designation, rasterisation, and compression. To identify protected areas relevant to the study region, the Ohsome API is used to reverse geocode the input extent and determine the corresponding countries. These country codes, together with the access token, protected area type (terrestrial or marine), and reference year, are then used to query the Protected Planet API.

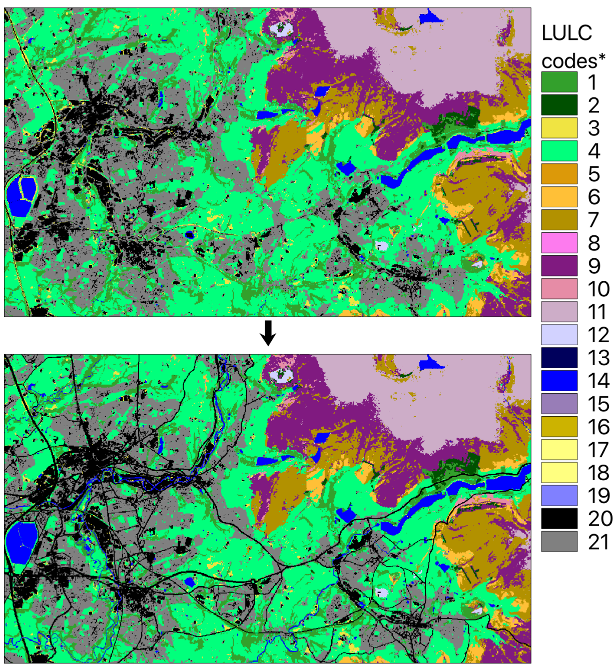
To account for the influence of protected areas on biodiversity and species movements, users can either specify a global resistance adjustment factor in the configuration file or assign dedicated resistance values to protected land cover classes through the reclassification table. The appropriate approach depends on the study objectives, desired level of detail, and characteristics of the LULC classification scheme. The resulting output consists of updated landscape resistance raster layers that incorporate the effects of protected areas

2.5. LULC Enrichment

The *enrich-lulc* command performs the following operations:

- Reads processing parameters and feature definitions from the configuration files.
- Loads the raster datasets and spatial features generated by *process-wdpa* and(or) *process-osm*.
- Applies buffers to linear features where necessary to ensure that features covering multiple raster cells are adequately represented.
- Rasterises the buffered geometries and merges them into a single raster layer.
- Creates a copy of the input raster dataset and updates it with the newly rasterised features.

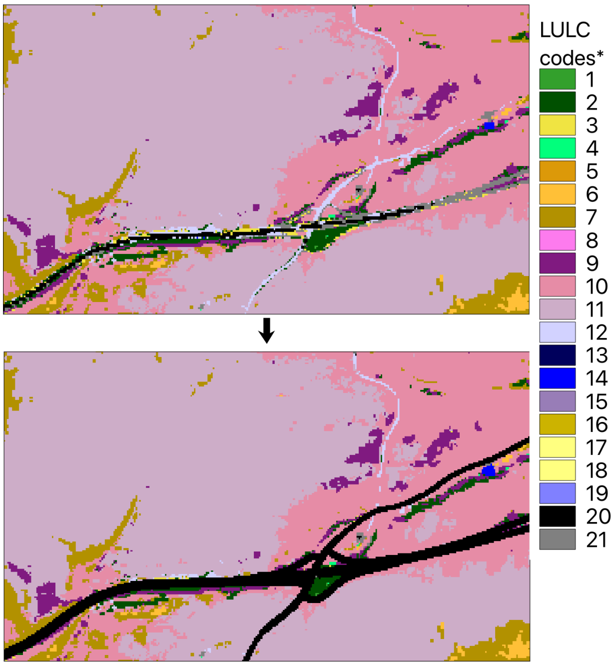
The output is an enriched LULC raster that preserves the original dataset specifications and classification scheme while incorporating the additional spatial information (Figure 4).



Includes data from © OpenStreetMap contributors (2026)

Figure 4. Non-enriched (top) and enriched (bottom) LULC maps (2023) for Northern England (*type codes are given according to Morton et al, 2024).

While the original LULC datasets may not consistently cover the road infrastructure, the enriched dataset encodes non-discrete linear features of roads and railways (Figure 5). The enrichment effect for LULC datasets directly affects further processing, such as habitat connectivity. As an illustration, indices such as Integrated Index of Connectivity or Betweenness Centrality, calculated via Graphab software, significantly differ for enriched and non-enriched datasets (Kriukov et al, 2025).



Includes data from © OpenStreetMap contributors (2026)

Figure 5. A local example of LULC datasets with roads fragmenting landscapes (Dark Peak region) with non-enriched version on top and enriched version on bottom, 2023

2.6. Ecological Enrichment and Landscape Resistance

The *recalc-impedance* command enables users to incorporate the influence of biodiversity stressors into the surrounding landscape through the modelling of edge effects which effectively increase landscape resistance. Multiple biodiversity stressors, such as roads or urban areas, can be considered in the configuration, arising either from particular LULC classes, or from OSM features. Edge effects represent the indirect impacts that these features can exert on surrounding habitats, including disturbance caused by human activity, noise, light, or pollution (Ewers and Didham, 2026).

To generate edge-effect layers, *recalc-impedance* performs the following steps:

- Encoding stressor data from the ecological configuration file;
- Rasterising stressors stemming from OSM data;
- Computing raster proximity to stressor pixels;
- Calculating edge effect, using type and decay multiplier for each stressor, which is defined by the user;
- Extracting maximum values of landscape resistance from all edge effects.

Sample ecological parameters are provided in the configuration directory; however, appropriate values are highly dependent on the species, ecosystem, and study objectives. Parameter selection may be informed by published literature, empirical observations, field surveys, camera-trap data, radio-tracking studies, or expert knowledge. In many ecological applications, the influence of stressors on landscape resistance is represented using an exponential decay function, reflecting the gradual reduction of impact with increasing distance from the source.

The ecological parameter configuration is provided in the YAML file and includes per-species or per-ecosystem landscape impedance values, decay rates, and habitat suitability multipliers for LULC classes (Table 3).

LULC	type	resistance	edge_effect	vector_refine
2	continental water	100	0	1
4	road infrastructure	200	1	1
6	urban	1000	1	0
14	bushes	1	0	0
15	coniferous forests	10	0	0
17	deciduous forests	10	0	0
18	burnt areas	50	0	0
22	vineyards	10	0	0

Table. 3. Fragment of the ecological configuration file for *Testudo hermanni* (Hermann's tortoise) in Catalonia. Column definitions: *LULC* – unique identifier of the land use/land cover class; *type* – full class name; *resistance* – inverse measure of landscape permeability to species movement; *edge_effect* – binary flag indicating whether the class acts as a biodiversity stressor; *vector_refine* – binary flag indicating whether the class is subject to vector-based enrichment.

Ecological configuration is highly flexible and should be configured by users on a case-by-case basis, considering available empirical information on species occurrences, previous studies or experiments (Beier et al, 2008).

2.7. Validation and Limitations

To date, Data4Land outputs have been validated against rasterised Ordnance Survey data products – Open Roads and Open Rivers (Kriukov et al, 2025). A result of 0.95 consistency for roads and 0.99 consistency for water features confirms the fitness-for-purpose of OSM as a source to enrich LULC datasets. Although recall values are low (0.48 for water features and 0.78 for roads), these discrepancies occur primarily among narrow and unpaved tracks, which present the least resistance to species movements and are not typically used in Data4Land enrichment.

The OSM data limitations relevant to Data4Land relate primarily to logical inconsistencies and feature omissions rather than geometric or topological errors, and are most pronounced at older timestamps. A full discussion is provided in Kriukov et al. (2025), but the key issues for users are summarised here.

OSM recording began in 2012, so LULC datasets predating this cannot be enriched from OSM alone and must rely on alternative vector sources. Even within the available record, earlier timestamps may underrepresent features: for example, the total digitised length of rivers and canals in Catalonia grew by 27.3% between 2012 and 2022, reflecting not real-world change but increasing user participation, improved geometric accuracy, and the progressive addition of previously unmapped features.

Beyond coverage, several tag-level inconsistencies affect data retrieval across time:

- Road classification tags are sometimes absent or inconsistently applied in 2012-2013 data, alongside with units of measurements.
- Road class designations occasionally change between timestamps without any corresponding change on the ground (e.g. a road reclassified from primary to secondary).
- Water feature geometry types have shifted over time: some rivers mapped as ways were later complemented with multipolygon representations, which affects feature counts and geometry handling, particularly in Catalonia at 2012, 2013 and 2017 timestamps.

Tag semantics have also evolved substantially across the OSM community. For water features specifically, the tag *waterway=riverbank* is now deprecated. In the UK, the convention for water reservoirs shifted from *landuse=reservoir* to n*atural=water* + *water=reservoir* around 2021, while the deprecated pair continued to be used in Catalonia well after this change. These differences mean that a single set of filtering rules cannot be applied uniformly across regions or time periods.

To handle this variability, Data4Land allows users to define custom combinations of OSM keys and values in the configuration file, adapting queries to their geographic context and the timestamp range under analysis. Users are strongly recommended to review tag history for their study area before finalising query definitions. The Ohsome API statistics and the OSM History Explorer provide useful starting points, and the query analysis scripts used for this study are available on GitHub.

2.8. Impact

Data4Land is being actively used to produce multiple raster datasets on terrestrial habitat connectivity for a wide range of forest-dwelling mammal species in Catalonia, e.g. *Martes martes*. It has been also used to produce habitat connectivity analysis at a more local scale in the Albera Natural Park for the reptile species *Testudo hermanni*, which is severely affected by conversion of natural habitats, mostly shrublands, into vineyards.

Although Data4Land has been mainly developed to analyse ecosystem structure and habitat connectivity more accurately, this tool can be potentially used in other applications. One application domain could be urban heat island (UHI) analysis. Impervious surfaces, above all road networks, absorb and re-emit solar radiation far more intensely than natural land cover, and their accurate representation in LULC data is critical for modelling land surface temperature (LST) dynamics and UHI intensity (Rao et al, 2023). However, roads are consistently underrepresented or entirely absent from medium-resolution satellite-derived LULC products, introducing systematic error into thermal analyses. By incorporating OSM road geometries with type-specific attributes (motorway, primary, secondary, tertiary, and so on), Data4Land enables the construction of more thematically complete LULC inputs for UHI studies, where the thermal contribution of different road categories can be modelled separately.

Another application is related to surface water monitoring and water quality assessment. The accurate delineation of water bodies, riparian corridors, and drainage networks in LULC datasets is a prerequisite for catchment-scale analyses, including riparian buffer assessment and the identification of potential pollution pathways from adjacent land uses (Li et al, 2021). Satellite-based products frequently miss narrow watercourses or misclassify water bodies obscured by riparian vegetation, while OSM provides well-maintained waterway and water body geometries validated by local contributors. Data4Land's water feature extraction allows users to supplement or correct existing LULC products with consistent hydrographic information, with direct relevance to water quality monitoring workflows.

For nature conservation applications, Data4Land provides a flexible framework for modelling landscape stressors beyond simple barrier/non-barrier classifications. Users can assign ecological weights to broad feature types or specific subtypes and define distinct impedance decay functions for different road categories. For example, motorways can be modelled as strong, short-range barriers, while minor roads can represent lower resistance and greater permeability, reflecting species-specific responses documented in the literature (Jacobson et al. 2016). Decay functions may be exponential or proportional, with user-defined distances of influence; this allows enriched resistance surfaces to be calibrated very flexibly on the basis of empirical data or expert knowledge.

3. Conclusions

The developed Data4Land tool is capable of performing a series of preprocessing and enrichment operations on land use and land cover (LULC) datasets:

1. Analysing input raster datasets to extract key metadata (CRS, NoData, bounding box, and spatial resolution).
2. Querying external REST APIs (Ohsome, Overpass Turbo, and Protected Planet) to retrieve vector patches used for LULC enrichment.
3. Harmonising retrieved data through value filtering, clipping to the study area, reprojection, and conversion from JSON/GeoJSON to GeoPackage.
4. Rasterising harmonised data, applying buffers where required, integrating rasterised features into the input LULC data.
5. Calculating edge effects associated with biodiversity stressors.

The workflow has been tested on land cover datasets covering 69,925.17 km^2 and 2,120 km^2 (corresponding to 8,981 × 8,651 and 5,466 × 3,881 pixels) in Catalonia and the UK. It is suitable for other study areas of similar or smaller extent and therefore has potential applicability for comparable projects worldwide. This scalable workflow enables automated habitat connectivity assessments for regional and local spatial planning, biodiversity conservation, and ecosystem services evaluations. It demonstrates the interoperability of four open-source technical components and highlights the need for multiple connectivity indices, as their temporal trends may diverge and are not always intercorrelated.

The Data4Land tool is capable of providing users with more accurate reflections of land use and land cover not only for nature conservation studies but also in other fields. The connectivity case study illustrates the value of a transparent and repeatable workflow in which parameters such as dispersal distance and edge effects of stressor features can be systematically varied in order to consistently assess their effect. The developed workflow is flexible and scalable and would be useful to implement at regional and local scales by planning authorities, environmental consultants, and nature conservation experts as a part of environmental impact assessments. We believe that the Data4Land tool will be a useful contribution to analysis not only of habitat connectivity, but of a range of landscape analyses where documentation of provenance and process are important.

Directions of future work may include:

- Providing the Data4Land package as a stand-alone pip library with version-stated dependencies.

- Combining Data4Land with machine learning models, e.g. segmentation workflows, as a pre-processing module.

- Developing additional functionality to add custom LULC changes, including positive changes, such as ecological corridors.

- Implement ingestion of other spatial features in raster format which may act opposed to biodiversity stressors, for example, small woody features.

- Implement AI agent for on-fly construction of queries to OSM database from the user input and follow-up enrichment of LULC datasets (supposedly through Natural Language Processing model).

- Implement custom data on protected areas in case a user doesn't have a token for Protected Planet API. This will help to bypass the limitations of the API and to use other data on protected areas, for example, small urban protected areas with relatively mild restrictions, not recorded in the World Database on Protected Areas.

- Prediction of habitat connectivity change under hypothetical LULC change, which is commonly referred as LULC scenario analysis (Gomes et al., 2020).

Acknowledgements

The Data4Land tool was developed during the AD4GD project. The AD4GD project is supported by Innovate UK, United Kingdom (grant number 10041654), the European Union, and Switzerland.

Data4Land GitHub:
https://github.com/AD4GD/pilot-2-preprocessing

The output data produced by Data4Land with the sample input data is accessible on Zenodo:
https://zenodo.org/records/15320266 (Dataset URL);
http://dx.doi.org/10.5281/zenodo.13991724 (Software URL).

References

Bastin, L., Kriukov, V., Lush, V., Serral, I., Hodson, T., Borger, C., Zamzow, M., Bruch, L., 2024: AD4GD D6.*1 Pilot Technical Implementation Planning, Implementation and Assessment*. 10.5281/zenodo.10839022

Beier, P., Majka, D.R., Spencer, W.D., 2008: Forks in the Road: Choices in Procedures for Designing Wildland Linkages. *Cons. Biol.*, 22, 836-851. doi.org/10.1111/j.1523-1739.2008.00942.x

Brennan, A., Bowman, J., Custode, L. et al., 2025: National-scale multispecies connectivity models represent movements for a majority of species tested. *Landsc. Ecol.*, 40, 200. doi.org/10.1007/s10980-025-02227-5

Cork, N.A., Fisher, R.S., Strong, N., Ferranti, E.J.S., Quinn, A.D., 2024: A systematic review of factors influencing habitat connectivity and biodiversity along road and rail routes in temperate zones. *Front. Environ. Sci.*, 12, 1369072. doi.org/10.3389/fenvs.2024.1369072

Foltête, J.-C., Vuidel, G., Savary, P., Clauzel, C., Sahraoui, Y., Girardet, X., Bourgeois, M., 2021: Graphab: An application for modeling and managing ecological habitat networks. *Software Impacts*, 8, 100065. doi.org/10.1016/j.simpa.2021.100065

Gomes, L.C., Bianchi, F.J.J.A., Cardoso, I.M., Schulte, R.P.O., Arts, B.J.M., Fernandes Filho, E.I., 2020: Land use and land cover scenarios: An interdisciplinary approach integrating local conditions and the global shared socioeconomic pathways. *Land Use Policy*, 97, 104723.
doi.org/10.1016/j.landusepol.2020.104723

González-Guerrero, O., Pons, X., 2020: The 2017 Land Use/Land Cover Map of Catalonia based on Sentinel-2 images and auxiliary data. *Revista de Teledetección*, 55, 81-92. doi.org/10.4995/raet.2020.13112

Haklay, M., Weber, P., 2008: OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, 7, 12-18. doi.org/10.1109/MPRV.2008.80

Jacobson, S.L., Bliss-Ketchum, L.L., de Rivera, C.E., Smith, W.P., 2016: A behavior-based framework for assessing barrier effects to wildlife from vehicle traffic volume. *Ecosphere*, 7(4), e01345. doi.org/10.1002/ecs2.1345

Keeley, A.T.H., Beier, P., Jenness, J.S., 2021: Connectivity metrics for conservation planning and monitoring. *Biol. Conserv.*, 255, 109008. doi.org/10.1016/j.biocon.2021.109008

Kriukov, V., Rahman, R., Bastin, L., 2025: Data4Land: How to enrich land use/land cover with historical vector(s). *SoftwareX*, 31, 102226. doi.org/10.1016/j.softx.2025.102226

Li, H., Zech, J., Ludwig, C., Fendrich, S., Shapiro, A., Schultz, M. et al., 2021: Automatic mapping of national surface water with OpenStreetMap and Sentinel-2 MSI data using deep learning. *Int. J. Appl. Earth Obs. Geoinf.*, 104, 102571. doi.org/10.1016/j.jag.2021.102571

Mahamune, S.S., Gavankar, N.L., 2025: A systematic review of LULC, LST and Urban Sprawl. *Int. Conf. on Future Technologies (ICFT), Sangli, India*, 1-8. doi.org/10.1109/ICFT66708.2025.11336382

Morton, R.D., Marston, C.G., O'Neil, A.W., Rowland, C.S., 2024: Land Cover Map 2020-2023 (10m classified pixels, GB).

NERC EDS Environmental Information Data Centre. doi.org/10.5285/7727ce7d-531e-4d77-b756-5cc59ff016bd

Patriarca, J., Fonte, C.C., Estima, J., de Almeida, J.-P., Cardoso, A., 2019: Automatic conversion of OSM data into LULC maps: comparing FOSS4G based approaches towards an enhanced performance. *Open Geospatial Data, Softw. Stand*., 4, 11. doi.org/10.1186/s40965-019-0070-2

Peak District National Park, 2021: The State of the Park Report. https://reports.peakdistrict.gov.uk/sotpr/ (15 June 2026).

Pechanec, V., Prokopová, M., Včeláková, R., Štěrbová, L., Cudlín, O., Pohanková, T., Grill, S., Salvati, L., Cudlín, P., 2025: Assessing habitat connectivity: A comparative analysis of four methodologies applied to peri-urban landscapes. *Sci. Total Environ*., 1003, 180729. doi.org/10.1016/j.scitotenv.2025.180729

Prima, M.-C., Renaud, J., Witté, I., Suarez, L., Rouveyrol, P., Fernando, M., Sacchi, A., Cosentino, F., Santini, L., Maiorano, L., Moreira, F., Dertien, J., Fernández, N., Thuiller, W., 2024: A comprehensive framework to assess multi-species landscape connectivity. *Methods Ecol. Evol*., 15, 2385-2399. doi.org/10.1111/2041-210X.14444

Radoux, J., Chomé, G., Jacques, D.C., Waldner, F., Bellemans, N., Matton, N., Lamarche, C., D'Andrimont, R., Defourny, P., 2016: Sentinel-2's Potential for Sub-Pixel Landscape Feature Detection. *Remote Sensing*, 8(6), 488. doi.org/10.3390/rs8060488

Rao, P., Tassinari, P., Torreggiani, D., 2023: Exploring the land-use urban heat island nexus under climate change conditions using machine learning approach: A spatio-temporal analysis of remotely sensed data. *Heliyon*, 9 (8), e18423. doi.org/10.1016/j.heliyon.2023.e18423

Sahraoui, Y., Foltête, J.-C., Clauzel, C., 2017: A multi-species approach for assessing the impact of land cover changes on landscape connectivity. *Landsc. Ecol*., 32, 1819-1835. doi.org/10.1007/s10980-017-0551-6

Schultz, M., Voss, J., Auer, M., Carter, S., Zipf, A., 2017: Open land cover from OpenStreetMap and remote sensing. *Int. J. Appl. Earth Obs. Geoinf.,* 63, 206-213. doi.org/10.1016/j.jag.2017.07.014

Torrás, O., Martín-Queller, E., Saura, S., 2009: Relating landscape structure, environment and management to biodiversity indicators estimated from forest inventory data in Catalonia (NE Spain). *Forest Systems*, 18(3), 322-337. doi.org/10.5424/fs/2009183-01072

UNEP-WCMC and IUCN, 2025: Protected Planet. The World Database on Protected and Conserved Areas (WDPCA). UNEP-WCMC and IUCN, Cambridge, UK. protectedplanet.net (15 June 2026).

Viana, C.M., Encalada, L., Rocha, J., 2019: The Value of OpenStreetMap Historical Contributions as a Source of Sampling Data for Multi-Temporal Land Use/Cover Maps. ISPRS *Int. J. Geo-Inf.*, 8(3), 116. doi.org/10.3390/ijgi8030116

Wang, Y., Sun, Y., Cao, X., Wang, Y., Zhang, W., Cheng, C., 2023: A review of regional and Global scale Land Use/Land Cover (LULC) mapping products generated from satellite remote sensing. *ISPRS J. Photogramm. Remote Sens*., 206, 311-334. doi.org/10.1016/j.isprsjprs.2023.11.014

Zeller, K.A., McGarigal, K., Whiteley, A.R., 2012: Estimating landscape resistance to movement: a review. *Landsc. Ecol*., 27, 777-797. doi.org/10.1007/s10980-012-9737-0