American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN: 2320-0936
Volume-15, Issue-8, pp-08-22
www.ajer.org
Research Paper Open Access

Spatio-Temporal Mining of Geo-Tagged Flickr Metadata for Tourism Hotspot Identification in Sudan: A Spatial Data Mining and Computational GIS Approach for Data-Scarce Contexts

Mohamed Gafar Elhag Elamin[1] and Dieter Fritsch[2]

[1]*College of Computer Science and Information Technology, Sudan University of Science and Technology, Khartoum, Sudan; m.gaafer.sd@gmail.com*

[2]*Institute for Photogrammetry and GeoInformation, University of Stuttgart, Geschwister-Scholl-Str. 24D, D-70174 Stuttgart, Germany; dieter.fritsch@ifp.uni-stuttgart.de*

Abstract

This paper develops and evaluates a reproducible computational GIS workflow for extracting spatial-temporal knowledge from geo-tagged Flickr metadata in Sudan, where official fine-scale tourism data are limited. Public metadata were acquired through a custom PHP collector using the Flickr API, stored in MySQL, and processed in Python and GIS environments. The final quality-control workflow recovered authentic Flickr user identifiers from photo URLs, filtered records against the Sudan national boundary, and reduced repeated-upload bias by retaining one observation for each photographer-location cell at approximately 100-110 m spatial resolution. From 19,699 raw regional records, 19,602 had valid capture dates and 8,162 were located within Sudan. The spatial analytical dataset contained 2,686 deduplicated photographer-location-cell observations contributed by 298 distinct Flickr users. DBSCAN with eps = 2 km and min_samples = 15 identified 20 candidate density clusters; 1,709 observations were assigned to clusters and 977 (36.4%) were classified as noise. The dominant structures remained the Khartoum metropolitan node, the Nile Valley archaeological corridor, and the Red Sea coastal zone after user-level deduplication. Independent spatial evidence supported this interpretation: the Clark-Evans nearest-neighbor index was 0.216 (z = -77.8), and 70 significant Getis-Ord Gi cells contained 84.1% of the analyzed activity. Fifteen of 20 clusters were within 5 km of a known geographic reference, while external Wikipedia/Wikimedia indicators showed a strong association between Flickr photographer diversity and Wikimedia Commons image availability (Spearman rho = 0.709, p < 0.001). Temporal patterns are reported for the deduplicated spatial unit and therefore represent the timing of retained digital visibility observations rather than tourist arrivals or independent visits. The findings demonstrate that transparent user identification, task-specific analytical units, parameter sensitivity analysis, and multi-source validation substantially improve the defensibility of social-media-derived tourism hotspot detection in data-scarce settings.*

Keywords: *spatial data mining; spatio-temporal mining; Flickr metadata; DBSCAN; computational GIS; tourism hotspots; Sudan; digital tourism visibility; user-generated geographic content; data-scarce contexts*

Date of Submission: 25-07-2026 Date of acceptance: 06-08-2026

I. Introduction

Recent advances in data mining, geospatial computing and Geographic Information Systems (GIS) have expanded the ability to derive knowledge from digital traces such as geo-tagged photographs, check-ins, online reviews and movement records. In tourism research, these sources are attractive because conventional statistics are commonly released at coarse administrative scales, after substantial delay, or without the location and time detail required for destination-level analysis (Salas-Olmedo et al., 2018; Tenkanen et al., 2017; Wood et al., 2013). The scientific challenge is not merely to display points on a map, but to construct a transparent workflow that transforms heterogeneous platform records into interpretable and reproducible spatial-temporal evidence.

This study treats tourism hotspot discovery as a knowledge discovery in databases problem. In the KDD formulation, pattern extraction is preceded by data selection, cleaning and transformation and followed by evaluation and interpretation (Fayyad et al., 1996). This distinction is important for geo-tagged social media because raw platform records can contain invalid coordinates, duplicate or near-duplicate uploads, inconsistent dates, uneven contributor activity and records outside the intended study area. A visually plausible hotspot map can therefore be methodologically weak if the analytical unit, user identity, clustering parameters and validation procedures are not explicitly defined.

Flickr provides a long historical archive of publicly shared photographs and exposes machine-readable metadata through its application programming interface. Available attributes may include latitude, longitude, capture time, upload time, photo identifier, user information, title and tags (Flickr, n.d.-a, n.d.-b). The distinction between capture and upload time is useful because historical images may be uploaded after they were taken. Flickr has consequently supported research on urban cores, visitor trajectories, protected-area use, attraction discovery and tourism mobility (Hollenstein & Purves, 2010; Lee & Tsou, 2018; Peng & Huang, 2017; Tenkanen et al., 2017).

Nevertheless, Flickr users do not constitute a probability sample of tourists. Platform participation is shaped by access, language, nationality, age, photographic practices, connectivity and the changing popularity of the service (Boyd & Crawford, 2012; Zook et al., 2017). Raw photo totals can also be dominated by a small number of prolific contributors. This paper therefore uses the concept of digital tourism visibility: spatial concentrations are interpreted as concentrations of geo-located photographic attention on Flickr, not as official visitor counts, tourist arrivals or a complete measure of destination demand.

Sudan is a compelling test case because it contains internationally significant urban, archaeological, riverine, desert, coastal and marine resources, while fine-scale tourism data are sparse. Important locations include the Khartoum metropolitan area; Meroe, Jebel Barkal, Nuri, El-Kurru, Naqa, Musawwarat es-Sufra, Kerma, Old Dongola and Soleb along the Nile Valley; and Port Sudan, Suakin and the Red Sea marine environment. UNESCO recognises the Archaeological Sites of the Island of Meroe, Jebel Barkal and the Sites of the Napatan Region, and the Sanganeb-Dungonab marine property (UNESCO World Heritage Centre, n.d.-a, n.d.-b, n.d.-c, n.d.-d). Yet tourism development is constrained by limited infrastructure, fragmented planning capacity and political instability (Mohamed et al., 2022).

The observed capture dates in the retained data extend from 2002 to 2019. Pre-platform capture dates are possible because Flickr allows older photographs to be uploaded later. The endpoint of 2019 provides a pre-war baseline before the profound mobility and security changes associated with Sudan's political transition and the armed conflict that began in April 2023 (Reuters, 2019, 2023). The baseline is not assumed to represent contemporary tourism conditions; rather, it preserves a historically coherent period for methodological evaluation.

The contribution is computational and empirical. Computationally, the study integrates API acquisition, relational storage, authentic Flickr user-ID recovery, boundary filtering, contributor-aware deduplication, DBSCAN clustering, parameter sensitivity testing, point-pattern statistics, GIS validation and Wikipedia/Wikimedia cross-validation. Empirically, it identifies persistent structures of Flickr-based digital visibility across Sudan. A central methodological improvement is the use of a deduplicated photographer-location-cell observation rather than an unqualified photo or visit count.

II. Research Problem, Gap and Objectives

Reliable spatial-temporal evidence is required to identify destination concentrations, compare regions, understand seasonality and support heritage and infrastructure decisions. In data-scarce contexts, however, official records may be incomplete, irregularly collected or aggregated beyond the scale at which tourism activity is experienced. User-generated geographic content offers supplementary evidence, but only if its platform biases and data-engineering problems are treated as part of the research problem rather than as minor preprocessing details.

Four gaps motivate the study. First, the geography of Flickr-based tourism research remains uneven, with most established applications located in data-rich cities, protected areas and mature tourism regions. Second, acquisition, database management, spatial filtering, user identification, clustering and GIS interpretation are often reported as separate steps rather than a traceable end-to-end workflow. Third, density clusters are frequently labelled as hotspots without adequate reporting of scale sensitivity, noise and contributor diversity. Fourth, independent validation remains weak in many social-media studies, especially where official comparison data are unavailable.

The aim is to identify and interpret spatial-temporal patterns of tourism-related digital visibility in Sudan through a reproducible computational GIS and spatial data mining workflow. The objectives are to: (1) acquire and structure public Flickr metadata; (2) validate coordinates, dates and authentic user identifiers; (3) create a

contributor-aware spatial analytical unit; (4) identify candidate density clusters using DBSCAN; (5) evaluate parameter sensitivity and spatial non-randomness; (6) describe regional and temporal variation; (7) validate clusters against geographic, heritage and independent digital reference sources; and (8) assess the usefulness and limitations of Flickr metadata in a data-scarce setting.
Table 1 translates these objectives into the corresponding computer-science task, processing operation and expected analytical output. It shows that hotspot detection is embedded within a broader knowledge-discovery workflow rather than treated as an isolated mapping exercise.

Table 1. Computational orientation of the revised study workflow.

Component	Computational focus	Primary output
Acquisition	Flickr API retrieval through a custom PHP collector	Raw regional metadata
Data management	MySQL storage, traceability and export	Structured records
Identity control	Flickr user ID parsed from photo URL	Authentic contributor identifier
Spatial filtering	Point-in-polygon and administrative spatial join	Sudan-only records and states
Deduplication	One photographer-location-cell observation	Contributor-aware spatial dataset
Spatial mining	DBSCAN and sensitivity testing	Candidate clusters and noise
Pattern evaluation	Clark-Evans NNI and Getis-Ord Gi*	Independent clustering evidence
Validation	Known sites, UNESCO and Wikipedia/Wikimedia indicators	Convergent geographic and digital evidence

III. Related Work and Research Positioning

3.1 Knowledge Discovery and Spatial Data Mining

Knowledge discovery provides the conceptual structure for transforming imperfect records into defensible evidence. Selection, preprocessing, transformation, mining and interpretation are analytically connected; errors in an early stage propagate into subsequent maps and statistical summaries (Fayyad et al., 1996). Spatial data mining adds location, distance, neighborhood and geographic context to this workflow. For point-event data, relevant structures include density concentrations, corridors, isolated observations, spatial inequality and departures from complete spatial randomness.

Volunteered geographic information research has emphasized both the analytical potential and variable quality of citizen-generated location data (Goodchild, 2007; Haklay, 2010; Elwood et al., 2012). These insights apply directly to social media: coordinate precision and volume do not guarantee representativeness, and the unit counted by an algorithm must be aligned with the research interpretation.

3.2 Geo-Tagged User-Generated Content in Tourism Analytics

Geo-tagged photographs can reveal visible places, tourism routes, temporal rhythms and destination preferences. Prior studies have used Flickr to describe city cores, infer trajectories, distinguish visitors and residents, monitor protected areas and compare social-media proxies with field or administrative sources (Derdouri & Osaragi, 2021; Domenech et al., 2020; Heikinheimo et al., 2017; Hollenstein & Purves, 2010; Martins & Santos, 2024). At broader scales, social media has also been used to quantify cultural and landscape values and to examine nature-based preferences (Hausmann et al., 2018; van Zanten et al., 2016).

These applications demonstrate that user-generated photographs can supplement scarce observations, but repeated uploads and highly active users can distort counts. Studies increasingly distinguish photos from users, user-days or other contributor-aware units. The present work extends that logic by recovering the authentic Flickr user identifier from the photo URL and creating one spatial observation for each user and approximately 100-110 m location cell.

3.3 Density-Based Clustering, Scale Sensitivity and Validation

DBSCAN is well suited to exploratory geographic clustering because it does not require the number of groups in advance, allows irregular cluster shapes and labels sparse observations as noise (Ester et al., 1996; Schubert et al., 2017). Its output is nevertheless controlled by the neighborhood radius, eps, and the minimum number of observations, min_samples. A single parameter pair cannot be treated as universally correct in a country-scale dataset with dense cities, linear heritage corridors and sparse desert areas.

This study therefore interprets DBSCAN clusters as scale-dependent candidates and evaluates them against a systematic parameter grid. It also separates algorithmic density from geographic meaning. Overall clustering is evaluated using the Clark-Evans nearest-neighbour index (Clark & Evans, 1954), while local high-value concentration is examined using Getis-Ord Gi* statistics (Getis & Ord, 1992; Ord & Getis, 1995). Known sites, UNESCO references and independent Wikipedia/Wikimedia indicators provide additional evidence rather than being treated as definitive ground truth.

3.4 Bias, Ethics and Digital Tourism Visibility

Social media data encode unequal access and platform-specific practices. Iconic, accessible and internationally visible locations may be overrepresented, while domestic tourism, low-connectivity regions and everyday recreation may be underrepresented. High platform visibility is therefore neither equivalent to tourism demand nor proof of tourism value.

Geo-tagged metadata also raise privacy and security concerns. The analysis is conducted at aggregate cell, cluster and regional levels; individual trajectories are not reported. Authentic user identifiers are required internally for deduplication and contributor counts, but should be hashed or withheld from any public analytical dataset. This is particularly important in a conflict-affected setting where detailed geospatial information may have unintended consequences (Zook et al., 2017).

IV. Data and Methods

The revised workflow was designed around a single authoritative analytical dataset so that every table, figure and statistical result could be traced to the same processing decisions. Figure 1 summarizes the sequence from API acquisition to uncertainty-aware interpretation.

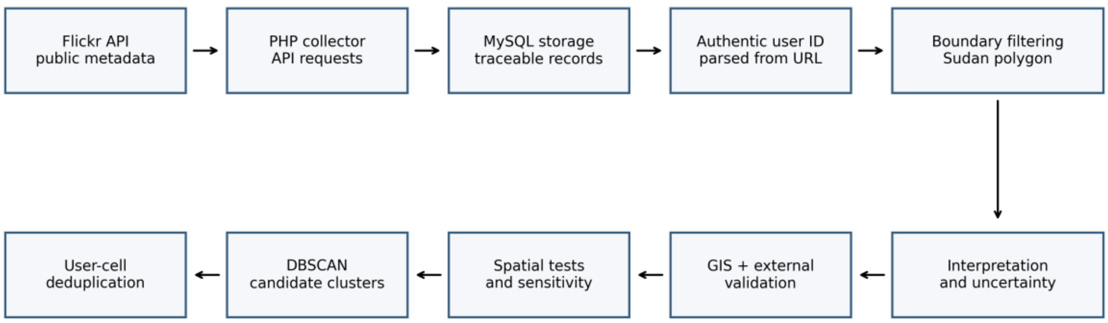

Figure 1. Revised computational GIS and spatial data mining workflow. The identity-control and contributor-aware deduplication stages are explicit components of the analytical pipeline.

4.1 Study Area and Temporal Scope

The study area is the current national territory of Sudan. National and first-order administrative boundaries were obtained from geoBoundaries, an openly licensed database designed for reproducible scientific analysis (Runfola et al., 2020). The geographic diversity of Sudan makes a uniform point distribution unlikely: urban access nodes, the Nile corridor, archaeological landscapes, Red Sea destinations and remote areas have markedly different levels of accessibility and digital documentation.

The retained capture dates span 2002-2019. Capture time, rather than upload time, was used where available. Because the final spatial analytical unit removes repeated user-cell combinations across the full period, temporal summaries in this paper describe the dates attached to retained spatial observations. They are not interpreted as independent trips, user-days or official tourist arrivals.

4.2 Data Acquisition and Relational Storage

Public Flickr metadata were collected using a custom PHP application that queried the Flickr API and stored results in a MySQL database. The collection preserved photo identifiers, coordinates, accuracy, capture time,

contextual place fields, URL and available owner-location metadata. Relational storage separated acquisition from analysis, supported duplicate checking and preserved traceability between API retrieval and downstream Python exports.
Python in Google Colab was used for data engineering, analysis and reproducible visualization, including pandas, NumPy, SciPy, scikit-learn, GeoPandas and Matplotlib (Harris et al., 2020; Hunter, 2007; McKinney, 2010; Pedregosa et al., 2011; Virtanen et al., 2020). QGIS supported geographic inspection and cartographic verification (QGIS Development Team, n.d.).

4.3 User Identification, Spatial Filtering and Analytical Unit

Latitude and longitude were converted to numeric values and records outside valid coordinate ranges were excluded. Capture times were parsed into year, month, day-of-week and hour components. Points were converted to WGS 84 spatial features and filtered by a point-in-polygon operation against the Sudan national boundary. A spatial join assigned first-order administrative units.
The authentic Flickr contributor identifier was parsed from the photo URL structure flickr.com/photos/<USER-ID>/<PHOTO-ID>. Records without a parseable user segment were excluded from user-based analysis. This correction is essential because an owner-place string represents a reported location, not a unique account, and a photo identifier represents an image rather than a photographer.
For the reference spatial model, coordinates were rounded to three decimal places, giving an approximate cell width of 100-110 m depending on latitude. One record was retained for each distinct photographer and rounded coordinate pair across the study period. The resulting unit is termed a deduplicated photographer-location-cell observation. It reduces repeated-upload bias while preserving multiple locations visited or photographed by the same contributor. It should not be labelled as an independent visit, tourist or raw photo count. As summarized in Table 2, the workflow reduced the regional collection from 19,699 raw records to 2,686 deduplicated photographer-location-cell observations contributed by 298 distinct Flickr users. And Figure 2 visualizes this reduction and makes clear that spatial filtering and contributor-aware deduplication were the two principal controls on the final analytical sample.

Table 2. Dataset reduction and final spatial analytical outputs.

Processing stage	Records/output	Interpretation
Raw API-derived regional records	19,699	Initial collection before temporal and spatial controls
Valid capture dates	19,602	Records usable for capture-date processing
Within Sudan boundary	8,162	Records retained after point-in-polygon filtering
Distinct Flickr users	298	Authentic user IDs recovered from photo URLs
Photographer-location-cell observations	2,686	Contributor-aware spatial analytical dataset
DBSCAN clustered observations	1,709	Observations assigned to 20 candidate clusters
DBSCAN noise	977 (36.4%)	Sparse observations not assigned to a cluster

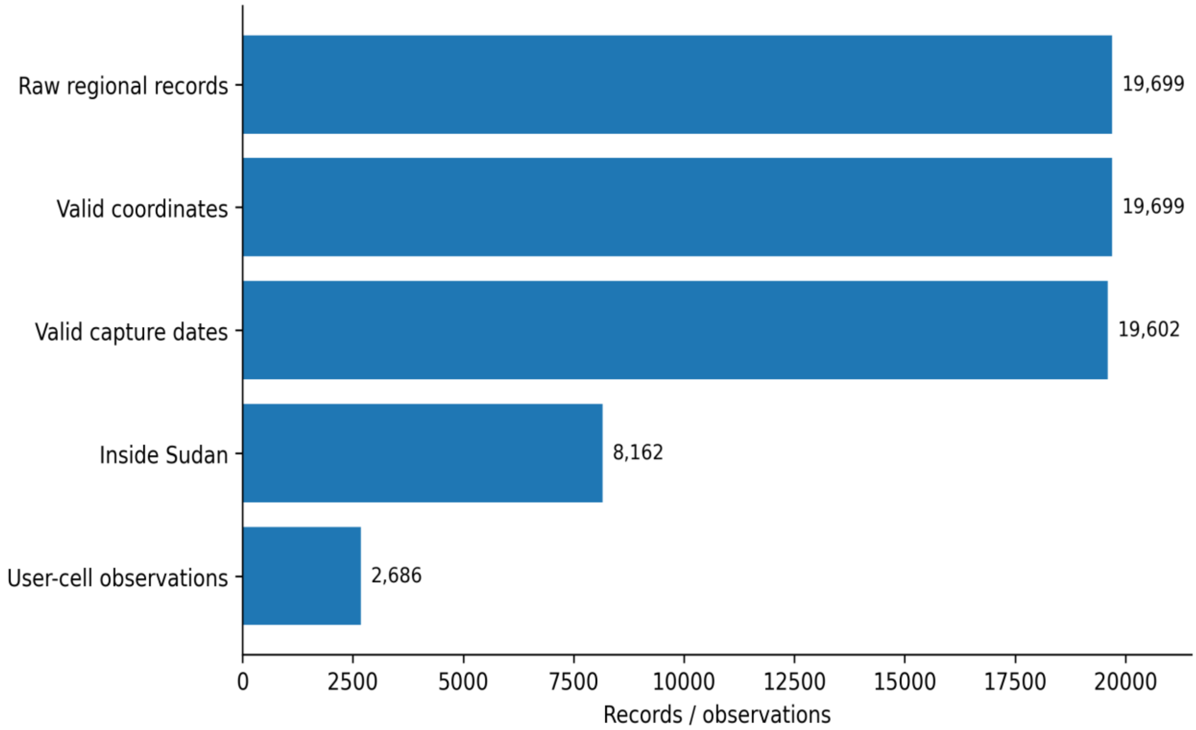

Figure 2. Data cleaning and filtering funnel from the raw regional collection to the final deduplicated photographer-location-cell dataset.

4.4 Temporal Analysis

Monthly and annual counts were calculated from the 2,686 retained photographer-location-cell observations. These summaries identify temporal structure in the final spatial evidence but deliberately avoid the language of visits. The Rayleigh test was used to evaluate directional concentration across months (Mardia & Jupp, 2000). Month-by-state association was evaluated using a chi-square test with Cramer's V, and differences in monthly distributions among major clusters were examined using the Kruskal-Wallis test (Kruskal & Wallis, 1952).

This task-specific interpretation is important. User-cell deduplication controls repeated spatial uploads but can merge repeated activity by the same user at the same cell across different dates. The results therefore provide a conservative temporal profile of retained digital visibility observations. Future work should construct a separate user-day or user-day-cell dataset when the primary objective is repeated visitation or trip timing.

4.5 DBSCAN Hotspot Detection and Sensitivity Analysis

DBSCAN was applied to the 2,686 spatial observations using the haversine metric. A reference configuration of eps = 2 km and min_samples = 15 was selected as a conservative local-scale neighborhood suitable for site-level concentrations. The 15-nearest-neighbour distance curve was used as a diagnostic, but the selected threshold was not claimed to arise from a single unambiguous elbow.

Robustness was evaluated across eps values of 1, 1.5, 2, 3 and 5 km and min_samples values of 5, 10, 15, 20 and 30. Across the 25 combinations, the number of clusters ranged from 10 to 60 and the noise proportion ranged from 15.4% to 57.9%. The reference configuration produced 20 clusters and 36.4% noise. This variation is reported explicitly because cluster counts are scale dependent.

DBSCAN defines density from observations rather than unique contributors. Each resulting cluster was therefore summarized by both the number of deduplicated observations and the number of distinct Flickr users. Clusters contributed by fewer than five users were retained as exploratory low-confidence candidates rather than confirmed tourism hotspots. Table 3 records the reference DBSCAN configuration and the interpretation assigned to core, border and noise observations, while Figure 3 complements the table by showing both the 15-nearest-neighbour diagnostic and the parameter-sensitivity results, demonstrating that the selected model is a scale-dependent but geographically interpretable solution.

Table 3. DBSCAN reference configuration and interpretation.

Element	Setting/result	Interpretive role
Analytical point	One photographer-location-cell observation	Controls repeated uploads at nearly identical coordinates
Distance metric	Haversine	Accounts for geographic coordinates on the Earth
eps	2 km	Reference neighbourhood radius
min_samples	15	Minimum local observation density
Reference output	20 clusters; 36.4% noise	Scale-dependent candidate structures
Sensitivity grid	25 parameter combinations	Clusters: 10-60; noise: 15.4-57.9%
User-diversity check	Distinct users per cluster	Supports confidence classification

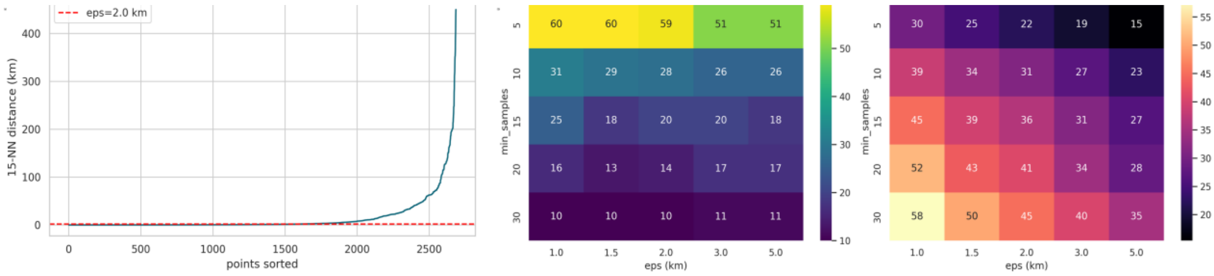

Figure 3. DBSCAN parameter assessment: (a) 15-nearest-neighbour distance diagnostic with the 2 km reference radius; and (b) sensitivity of cluster count and noise percentage to alternative eps and min_samples values.

4.6 Spatial Pattern Evaluation

The Clark-Evans nearest-neighbor index compared the observed mean nearest-neighbor distance with the expectation under complete spatial randomness within the Sudan boundary (Clark & Evans, 1954). A value below one indicates clustering. Polygon area was computed in an equal-area projection before the expected distance and z-score were calculated.

Local high-activity concentration was evaluated using Getis-Ord Gi* (Getis & Ord, 1992; Ord & Getis, 1995). Observations were aggregated to a 0.5-degree national grid, transformed using log(1+n), and evaluated with a binary distance-band neighbourhood. Significant positive cells were treated as a coarse national-scale corroboration of concentration rather than precise site boundaries.

4.7 Geographic and External Digital Validation

Candidate cluster centroids were compared with 69 known tourism, urban and heritage references assembled from OpenStreetMap and Wikidata (OpenStreetMap contributors, n.d.; Wikidata contributors, n.d.). The nearest-reference distance was calculated for each cluster. Separate UNESCO reference coordinates represented Meroe, Jebel Barkal/Napatan sites, Sanganeb Marine National Park and Dungonab Bay-Mukkawar Island; a 15 km proximity threshold was used only as an interpretive overlay, not as proof of site identity.

A second form of convergent validation used selected Wikipedia and Wikimedia indicators for the 20 clusters: pageviews during 2015-2019, language-edition count, article length, article image count, Wikidata sitelinks and Wikimedia Commons images. Associations with Flickr cluster popularity were evaluated using Spearman rank correlation (Spearman, 1904). These indicators measure independent digital visibility and are not assumed to be tourism statistics. The Wikimedia Analytics and MediaWiki APIs supplied the underlying metrics (Wikimedia Foundation, n.d.).

V. Results

5.1 Data Quality and Analytical Dataset

The workflow reduced 19,699 raw regional records to 19,602 records with valid capture dates and 8,162 points within Sudan. Authentic user IDs recovered from Flickr URLs represented 298 distinct contributors. User-cell deduplication produced 2,686 spatial analytical observations, a reduction of 67.1% from the in-Sudan records.

The reduction is an analytical result in its own right: raw photographic volume was strongly affected by repeated spatial contributions, and hotspot estimates based only on photos would have overstated some locations.

The final dataset represented 19 first-order administrative units. Its observations are internally consistent across the tables, figures and tests presented below. The 2,686 count is used only for the spatial analytical unit; the paper avoids equating it with visits or tourists.

5.2 Spatial Distribution and Regional Concentration

The spatial observations were highly uneven. Khartoum contained 946 observations (35.2%), followed by Northern State with 914 (34.0%), River Nile with 429 (16.0%) and Red Sea with 116 (4.3%). Together these four states accounted for 2,405 observations, or 89.5% of the spatial analytical dataset. The corrected user-aware ranking therefore places Khartoum slightly ahead of Northern State, while still confirming the major importance of the Nile Valley archaeological landscape.

The point pattern forms three interpretable systems: a dense metropolitan concentration around Khartoum, a north-south chain of observations along the Nile and archaeological corridor, and a smaller coastal signal around Port Sudan and Suakin. Western, southern and desert regions contain scattered records. Low platform visibility in these areas cannot be interpreted as absence of tourism potential because accessibility, conflict, connectivity and Flickr adoption vary substantially.

Figure 4 jointly displays the national point pattern and the state-level distribution. Its two panels show that the strongest concentration is shared between Khartoum and Northern State and that the Nile corridor dominates the wider spatial structure.

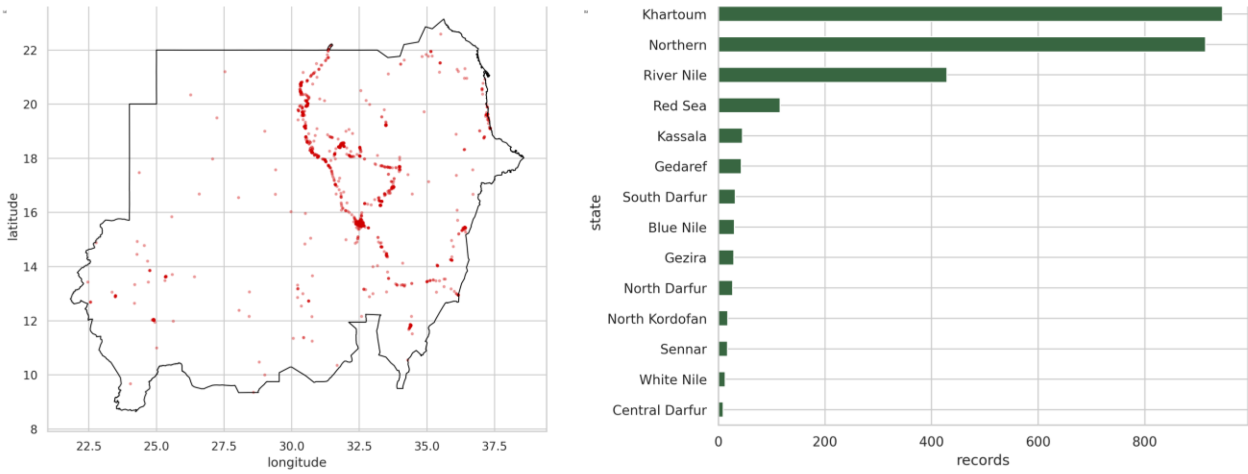

Figure 4. Spatial distribution of the final analytical unit: (a) 2,686 deduplicated photographer-location-cell observations within Sudan; and (b) observations by first-order administrative unit.

5.3 Temporal Structure of the Retained Spatial Observations

The monthly distribution of the retained user-cell observations was concentrated in the cooler part of the year. January was highest with 599 observations, followed by December with 473, April with 283 and November with 272. May was lowest with 59 observations and July contained 62. The Rayleigh result confirmed significant monthly concentration ($R = 0.387$, $z = 403.0$, $p < 0.001$).

Annual counts were low in the earliest years, increased after 2010 and reached their highest retained count in 2016 (528). The remaining high years were 2013 (369), 2017 (298), 2011 (281) and 2018 (267). These values differ from raw-photo peaks because one user-cell combination is retained only once. Accordingly, the annual pattern should be read as the capture-date distribution of spatially distinct contributor-cell evidence, not a direct tourism time series.

Month and state were significantly associated, chi-square (77, $n = 2{,}555$) = 1,210.6, $p < 0.001$, with Cramer's $V = 0.260$. Monthly distributions also differed among major clusters (Kruskal-Wallis $H = 34.5$, $p < 0.001$), indicating that seasonality was geographically differentiated.

Figure 5 presents the monthly and annual distributions of the retained spatial observations. The figure confirms concentration in the cooler part of the year and a marked post-2010 increase, while its interpretation remains limited to deduplicated user-cell records rather than independent visits.

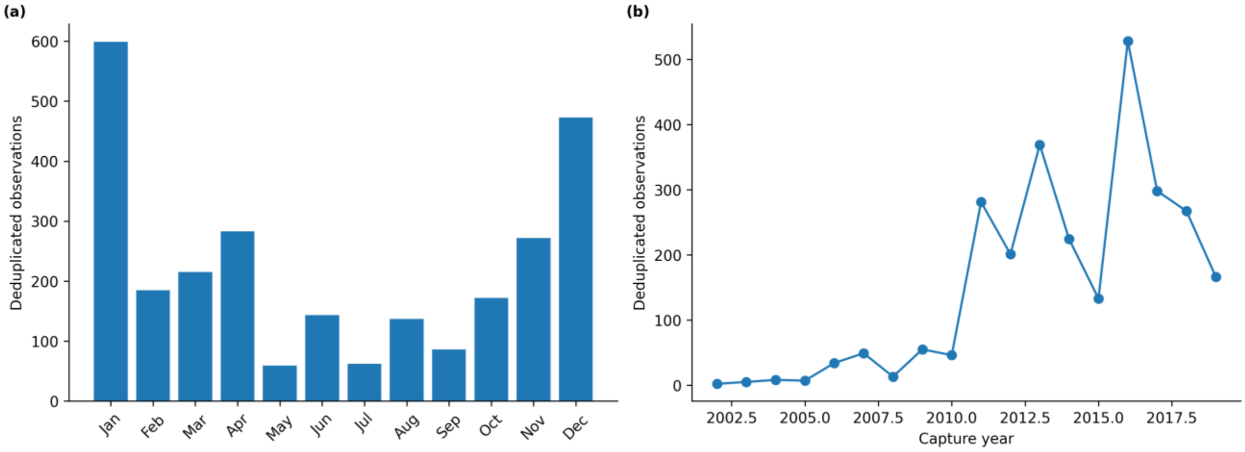

Figure 5. Temporal distribution of the deduplicated photographer-location-cell observations: (a) monthly counts and (b) annual counts by capture date. These counts describe the retained spatial unit, not independent visits.

5.4 DBSCAN-Derived Candidate Hotspots

The reference DBSCAN model identified 20 candidate density clusters. A total of 1,709 observations were assigned to clusters, while 977 observations (36.4%) remained noise. Noise is analytically meaningful because it represents isolated or low-density digital traces that did not satisfy the selected local threshold.

The largest cluster was the Khartoum metropolitan area, containing 796 observations from 185 distinct users. The next largest user-diverse clusters were Kabushiyah/Meroe and Marawi/Jebel Barkal, each contributed by 33 users. Old Dongola, Nuri, El-Kurru, Naqa, Musawwarat es-Sufra, Soleb, Kerma and related locations reinforce the Nile Valley heritage corridor. Port Sudan and Suakin provide the principal coastal signal.

Twelve clusters were contributed by at least ten distinct users, six by five to nine users, and two by fewer than five users. The Nyala and Kurgus-area clusters were therefore classified as low-confidence exploratory candidates. Their inclusion as algorithmic outputs is transparent, but they should not be presented as confirmed tourism hotspots without additional evidence.

Table 4 compares the major clusters using both observation counts and distinct-user diversity, allowing low-contributor clusters to be distinguished from more defensible candidate hotspots.

Figure 6 maps all 20 clusters together with the surrounding noise and shows the dominance of Khartoum and the chain of heritage clusters along the Nile.

Table 4. Major user-diverse DBSCAN candidate clusters.

Rank	Cluster/locality	State	Observations	Users	Interpretation
1	Khartoum metropolitan area	Khartoum	796	185	Dominant urban visibility node
2	Kabushiyah / Meroe	River Nile	189	33	Pyramid and archaeological landscape
3	Marawi / Jebel Barkal	Northern	172	33	Napatan heritage landscape
4	Tangasi / El-Kurru	Northern	46	21	Royal cemetery heritage cluster
5	Musawwarat area	River Nile	50	19	Archaeological cluster
6	Naqa area	River Nile	54	18	Archaeological cluster
7	Old Dongola	Northern	86	17	Christian/Nile Valley heritage cluster
8	Nuri	Northern	53	16	Napatan royal cemetery cluster
9	Wawa / Soleb	Northern	34	12	Temple heritage cluster
10	Dongola / Tabo area	Northern	16	11	Northern Nile cluster
11	Port Sudan	Red Sea	28	10	Coastal urban cluster

Rank	Cluster/locality	State	Observations	Users	Interpretation
12	Dongola / Kerma area	Northern	25	10	Ancient Nubian heritage cluster

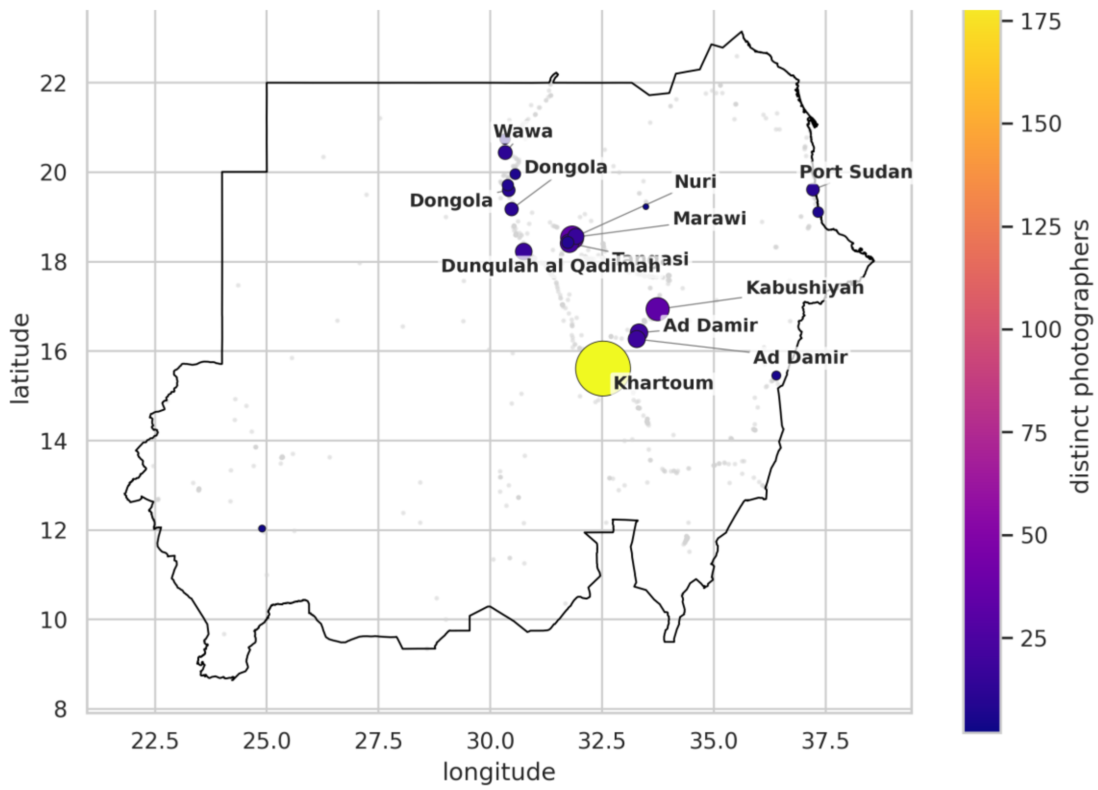

Figure 6. Spatial distribution of the 20 candidate DBSCAN clusters identified from 2,686 deduplicated photographer-location-cell observations. Symbol size and colour represent distinct photographer diversity; grey points are noise.

5.5 Spatial Corroboration and Geographic Validation

The Clark-Evans index strongly rejected a random national point pattern. The observed index was 0.216 with z = -77.8, indicating that the mean nearest-neighbour distance was far smaller than expected under complete spatial randomness. This result supports the existence of national-scale concentration independently of DBSCAN. Figure 7 visualises the two independent spatial corroboration tests: the Getis-Ord Gi* concentration grid and the Clark-Evans comparison with random expectation. Both panels indicate that the national distribution is strongly clustered.

The Getis-Ord Gi* analysis detected 70 significant positive grid cells. These cells contained 84.1% of the analytical observations. The concentration followed the Khartoum-Nile axis and parts of the eastern corridor, providing a coarse regional corroboration of the density structures. Because the grid cell size was 0.5 degrees, the result is interpreted at national and regional scales rather than as a boundary for individual attractions.

Geographic validation further supported the major clusters. The median distance from a cluster centroid to the nearest known reference was 1.02 km, and 15 of 20 clusters were within 5 km of a known OSM or Wikidata location. Four clusters fell within 15 km of the selected UNESCO reference coordinates. Proximity was strongest for Meroe, Jebel Barkal, El-Kurru, Nuri, Naqa, Musawwarat, Old Dongola, Soleb and Kerma. More distant or poorly matched clusters were treated cautiously because reference databases are incomplete and some place labels are broad. Figure 8 places the cluster centroids against OSM/Wikidata and UNESCO reference locations. The close spatial correspondence of most major clusters provides contextual support while preserving the distinction between proximity and definitive site classification.

Cluster contributor diversity declined significantly with distance from Khartoum ($r = -0.717$, $p < 0.001$). This relationship is consistent with a core-periphery pattern in digital visibility, but it may combine tourism, access, population, transport and platform effects.

Table 5 consolidates the statistical and geographic validation evidence and shows that the DBSCAN pattern is supported by multiple independent indicators rather than by a single clustering algorithm, while Figure 7 represents the independent spatial confirmation and Figure 8 the geographic validation.

Table 5. Statistical and validation evidence for spatial structure.

Test/indicator	Final result	Interpretation
Clark-Evans NNI	0.216; z = -77.8	Strong clustering relative to spatial randomness
Getis-Ord Gi*	70 significant hot cells	84.1% of activity in significant positive cells
DBSCAN reference model	20 clusters; 36.4% noise	Conservative local-scale candidate structures
Nearest known site	Median = 1.02 km	15 of 20 clusters within 5 km
UNESCO overlay	4 clusters within 15 km	Proximity to selected World Heritage references
Distance from Khartoum	r = -0.717; p < 0.001	Contributor diversity declines with distance
Site-popularity Gini	0.693	Substantial concentration among visible places
Photographer-activity Gini	0.836	Raw contribution activity strongly unequal

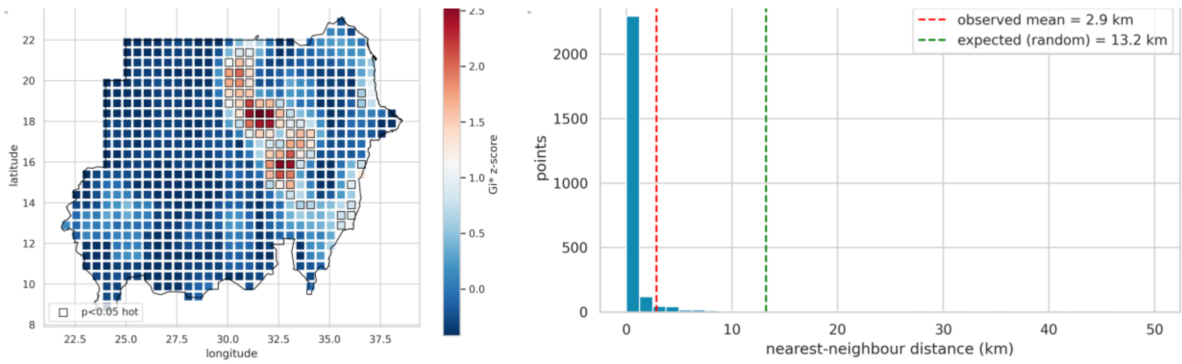

Figure 7. Independent spatial corroboration: (a) national-scale Getis-Ord Gi* grid, with outlined cells indicating significant positive concentration; and (b) observed and random-expected nearest-neighbour distances used in the Clark-Evans index.

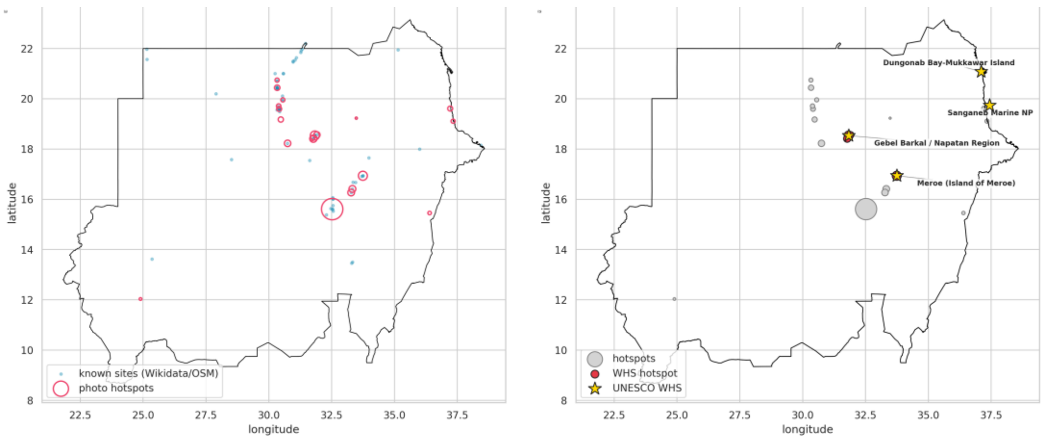

Figure 8. Geographic validation of candidate clusters: (a) cluster centroids compared with known OSM/Wikidata references; and (b) cluster proximity to selected UNESCO World Heritage coordinates. The overlays provide convergent context rather than definitive cluster labels.

5.6 Inequality and External Digital Convergence

The Gini coefficient for site popularity was 0.693, confirming that a limited number of places accounted for a large share of Flickr visibility. Photographer activity was even more unequal, with Gini = 0.836. This contrast provides direct justification for contributor-aware deduplication: without controlling repeated users, raw photo counts would partly measure uploader intensity rather than place visibility.

External digital indicators provided mixed but informative convergence. Flickr photographer diversity had a strong positive association with Wikimedia Commons image availability (Spearman rho = 0.709, p = 0.0005, n = 20) and a moderate significant association with Wikipedia article image count (rho = 0.502, p = 0.0285, n = 19). The association with 2015-2019 Wikipedia pageviews was moderate but narrowly missed the conventional significance threshold (rho = 0.458, p = 0.0562, n = 18). Language editions and Wikidata sitelinks were not significantly related to Flickr popularity.

These results do not establish Wikipedia as ground truth. They suggest that visually documented places on Flickr also tend to have richer image representation in other open digital ecosystems. The stronger performance of image-based indicators than pageviews or sitelinks is conceptually plausible because Flickr primarily records photographic attention. Figure 9 compares Flickr contributor diversity with four external digital indicators. The strongest relationships are associated with image-based measures, particularly Wikimedia Commons availability, whereas pageviews and sitelinks provide weaker evidence.

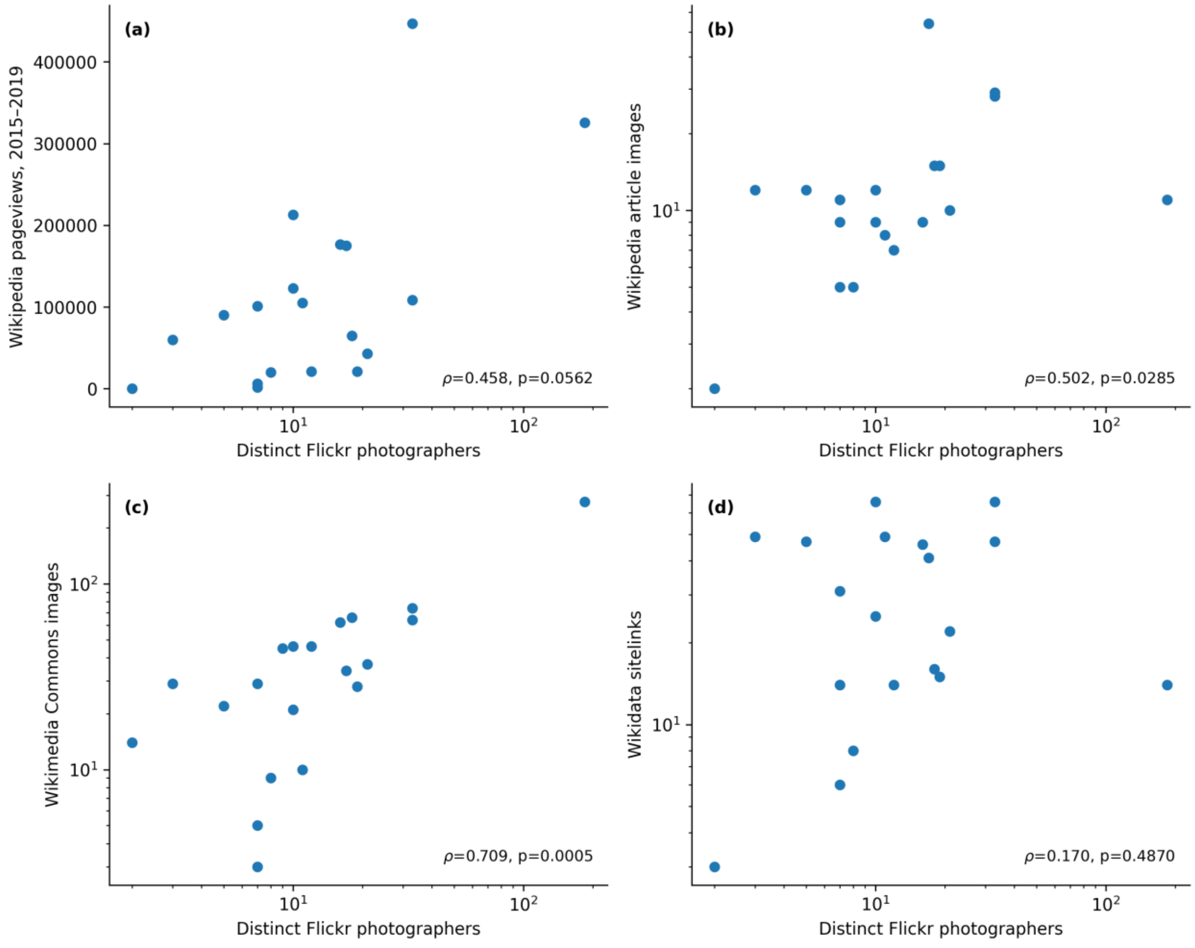

Figure 9. Convergent digital validation between distinct Flickr photographers and external indicators: (a) Wikipedia pageviews, (b) Wikipedia article images, (c) Wikimedia Commons images, and (d) Wikidata sitelinks. Axes are logarithmic where values are positive; rho denotes Spearman rank correlation.

VI. Discussion

The revised analysis demonstrates that methodological control changes the magnitude of the results without overturning their principal geography. After authentic user-ID recovery and user-cell deduplication, the analytical dataset and number of DBSCAN clusters were substantially smaller than in photo-based models.

Nevertheless, Khartoum, the Nile Valley archaeological corridor and the Red Sea coast remained the dominant spatial structures. Their stability across a stricter analytical unit suggests that they are not solely artefacts of repeated uploads.

The Khartoum cluster reflects a combination of urban visibility, access, administration, museums, transport and arrival functions. Its 185 distinct contributors and 796 deduplicated observations make it the most defensible cluster in the dataset. The Nile Valley clusters are individually smaller but geographically coherent and closely aligned with archaeological references. Together they form a distributed heritage system rather than a single circular hotspot. This illustrates why density-based clustering is useful: it reveals multiple irregularly spaced site concentrations without requiring a predefined number of groups.

The agreement among DBSCAN, Clark-Evans, Getis-Ord Gi*, reference-site proximity and open digital image indicators strengthens the interpretation of major clusters as meaningful digital visibility structures. No single method is decisive. DBSCAN is parameter dependent; nearest-neighbor analysis tests overall pattern rather than attraction identity; Gi* depends on grid and neighborhood scale; reference databases are incomplete; and Wikipedia/Wikimedia measure a different form of online attention. Convergence across these imperfect sources is therefore more persuasive than any one output.

The sensitivity analysis is especially important for reviewer-facing transparency. Alternative parameter choices produced between 10 and 60 clusters and markedly different noise rates. The reference model should consequently be interpreted as a conservative, site-level partition rather than the unique true clustering of tourism in Sudan. In a highly heterogeneous national dataset, future work could compare HDBSCAN or OPTICS, regionalize the analysis, or model multiple density scales (Schubert et al., 2017).

The temporal result requires equal caution. January and December dominate the retained user-cell observations, which is consistent with cooler travel conditions, but the unit intentionally removes repeated user-cell combinations across time. It is therefore unsuitable for estimating visit frequency. The temporal component shows when the retained spatial evidence was captured; a dedicated user-day model is required before making stronger claims about repeated visitation, trip volume or annual tourism trends.

The study contributes to computer science and information technology by making data identity, analytical unit and validation explicit parts of a spatial KDD workflow. Tourism is the application domain, while the methodological contribution is a reproducible route from noisy platform metadata to conservative and uncertainty-aware spatial evidence. The workflow is transferable to other data-scarce contexts, provided that platform-specific identifiers, boundaries and validation sources are carefully adapted.

VII. Limitations and Ethical Considerations

Flickr users are not representative of all visitors. International travelers, photographers, researchers and digitally connected contributors may be overrepresented, while domestic tourism and low-connectivity communities may be underrepresented. Platform popularity also changed over the study period. The results therefore measure Flickr-based digital visibility, not tourism demand.

The photographer-location-cell unit controls repeated uploads but merges repeated activity by the same user at the same cell across dates. This makes the spatial model conservative but limits temporal inference. Conversely, DBSCAN may still form a cluster from a small number of users occupying multiple neighboring cells. Distinct-user diversity was therefore used as a confidence indicator, and clusters with fewer than five users were explicitly labelled low confidence.

Coordinate rounding imposes a spatial scale, reverse-geocoded locality labels may not match tourism-site names, and the common 2 km DBSCAN radius cannot fully accommodate national density heterogeneity. Getis-Ord analysis used a coarse 0.5-degree grid. The known-site inventory and Wikipedia/Wikimedia mappings are incomplete and may contain broad or imperfect place matches. Proximity and digital correlation support interpretation but do not prove that all records were generated by tourists.

Only publicly shared metadata were analysed. Individual histories are not reported, and results are aggregated. Authentic Flickr identifiers are used internally for deduplication; they should be hashed or removed before any public release. Post-2019 data should be analysed separately because conflict and displacement change both mobility and the ethical meaning of geo-tagged traces.

VIII. Conclusion and Future Work

This paper presented a revised computational GIS workflow for mining geo-tagged Flickr metadata in Sudan. From 19,699 raw regional records, 19,602 had valid capture dates and 8,162 were spatially located within Sudan. Authentic user-ID recovery and deduplication at the photographer-location-cell level produced 2,686 spatial observations contributed by 298 users. DBSCAN identified 20 candidate clusters, with 1,709 clustered observations and 977 noise observations.

The main geographic conclusion remained stable after the stricter user-aware processing: digital tourism visibility was concentrated in the Khartoum metropolitan node, the Nile Valley archaeological corridor and the Red Sea coastal zone. The conclusion was supported by a strongly clustered Clark-Evans index, significant Getis-Ord cells, proximity to known sites and UNESCO references, and selected Wikipedia/Wikimedia image indicators. These findings show that social-media metadata can provide useful exploratory evidence in a data-scarce setting when the unit of analysis and uncertainty are handled explicitly.

Future research should construct a dedicated user-day or user-day-cell temporal dataset; compare DBSCAN with HDBSCAN, OPTICS and multi-scale clustering; incorporate image-content classification; expand and manually verify tourism-site inventories; integrate official records and field validation; and treat the post-2019 period as a separate conflict-sensitive study. A composite confidence score combining cluster density, distinct-user diversity, parameter stability and geographic validation would also improve prioritisation of candidate hotspots.

Acknowledgements

The manuscript was prepared as part of a doctoral research project on computational GIS and spatial-temporal mining of user-generated geospatial data for tourism hotspot identification in Sudan, Sudan University of Science and Technology, supervised by the second author.

Data Availability Statement

The study uses publicly available Flickr metadata collected through the Flickr API. Aggregated outputs, processing code and a privacy-preserving version of the analytical data may be made available upon reasonable request, subject to Flickr terms, ethical considerations and removal or hashing of user identifiers.

Ethics and Responsible Data Use Statement

The analysis uses public metadata and reports only aggregate spatial patterns. It does not disclose personal identities or reconstruct individual travel histories. Authentic account identifiers were used internally only to control duplicate contributions and should not be distributed in identifiable form.

Conflicts of interest**.**

The authors declare no known conflicts of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

References

[1]. boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

[2]. Clark, P. J., & Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology, 35(4), 445-453. https://doi.org/10.2307/1931034

[3]. Derdouri, A., & Osaragi, T. (2021). A machine learning-based approach for classifying tourists and locals using geotagged photos: The case of Tokyo. Information Technology & Tourism, 23(4), 575-609. https://doi.org/10.1007/s40558-021-00208-3

[4]. Domenech, A., Mohino, I., & Moya-Gomez, B. (2020). Using Flickr geotagged photos to estimate visitor trajectories in World Heritage cities. ISPRS International Journal of Geo-Information, 9(11), 646. https://doi.org/10.3390/ijgi9110646

[5]. Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. Annals of the Association of American Geographers, 102(3), 571-590. https://doi.org/10.1080/00045608.2011.595657

[6]. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226-231). AAAI Press.

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

[8]. Flickr (n.d.). Flickr API: flickr.photos.getInfo. https://www.flickr.com/services/api/flickr.photos.getInfo.html

[9]. Flickr (n.d.). Flickr API: flickr.photos.search. https://www.flickr.com/services/api/flickr.photos.search.html

[10]. GeoPandas developers (2020). GeoPandas: Python tools for geographic data. https://doi.org/10.5281/zenodo.3946761

[11]. Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. Geographical Analysis, 24(3), 189-206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

[12]. Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. GeoJournal, 69(4), 211-221. https://doi.org/10.1007/s10708-007-9111-y

[13]. Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B: Planning and Design, 37(4), 682-703. https://doi.org/10.1068/b35097

[14]. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., & Smith, N. J. (2020). Array programming with NumPy. Nature, 585, 357-362. https://doi.org/10.1038/s41586-020-2649-2

[15]. Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., & Di Minin, E. (2018). Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. Conservation Letters, 11(1), e12343. https://doi.org/10.1111/conl.12343

[16]. Heikinheimo, V., Di Minin, E., Tenkanen, H., Hausmann, A., Erkkonen, J., & Toivonen, T. (2017). User-generated geographic information for visitor monitoring in a national park: A comparison of social media data and visitor survey. ISPRS International Journal of Geo-Information, 6(3), 85. https://doi.org/10.3390/ijgi6030085

[17]. Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr to describe city cores. Journal of Spatial Information Science, 1(1), 21-48. https://doi.org/10.5311/JOSIS.2010.1.3

[18]. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

[19]. Jiang, W., Xiong, Z., Su, Q., Long, Y., Song, X., & Sun, P. (2021). Using geotagged social media data to explore sentiment changes in tourist flow: A spatiotemporal analytical framework. ISPRS International Journal of Geo-Information, 10(3), 135. https://doi.org/10.3390/ijgi10030135

[20]. Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583-621. https://doi.org/10.1080/01621459.1952.10483441

[21]. Lee, J. Y., & Tsou, M. H. (2018). Mapping spatiotemporal tourist behaviors and hotspots through location-based photo-sharing service (Flickr) data. In Progress in Location Based Services 2018 (pp. 315-334). Springer. https://doi.org/10.1007/978-3-319-71470-7_16

[22]. Liu, L., Zhang, Y., Ma, Z., & Wang, H. (2023). An analysis on the spatiotemporal behavior of inbound tourists in Jiaodong Peninsula based on Flickr geotagged photos. International Journal of Applied Earth Observation and Geoinformation, 120, 103349. https://doi.org/10.1016/j.jag.2023.103349

[23]. Mardia, K. V., & Jupp, P. E. (2000). Directional Statistics. Wiley, Chichester, UK.

[24]. Martins, M., & Santos, A. (2024). Exploring the potential of Flickr user-generated content for tourism research: Insights from Portugal. European Journal of Tourism, Hospitality and Recreation, 14(2), 258-272. https://doi.org/10.2478/ejthr-2024-0019

[25]. McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 56-61). https://doi.org/10.25080/Majora-92bf1922-00a

[26]. Mohamed, A., Birsen, A. G., & Bilim, Y. (2022). Tourism development potentials and challenges in Sudan. Journal of Economic Cooperation and Development, 43(4), 23-44.

[27]. OpenStreetMap contributors (n.d.). OpenStreetMap. https://www.openstreetmap.org/

[28]. Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. Geographical Analysis, 27(4), 286-306. https://doi.org/10.1111/j.1538-4632.1995.tb00912.x

[29]. Payntar, N. D., Hsiao, W. L., Covey, R. A., & Grauman, K. (2020). Learning patterns of tourist movement and photography from geotagged photos at archaeological heritage sites in Cuzco, Peru. arXiv. https://arxiv.org/abs/2006.16424

[30]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[31]. Peng, X., & Huang, Z. (2017). A novel popular tourist attraction discovering approach based on geo-tagged social media big data. ISPRS International Journal of Geo-Information, 6(7), 216. https://doi.org/10.3390/ijgi6070216

[32]. QGIS Development Team (n.d.). QGIS User Guide. https://docs.qgis.org/latest/en/docs/user_manual/

[33]. Reuters (2019-07-03). Abandoned by the UAE, Sudan's Bashir was destined to fall. https://www.reuters.com/investigates/special-report/sudan-bashir-fall/

[34]. Reuters (2023-04-17). What's behind the war in Sudan?. https://www.reuters.com/world/africa/whats-behind-sudans-crisis-2023-04-17/

[35]. Runfola, D., Anderson, A., Baier, H., Crittenden, M., Dowker, E., Fuhrig, S., Goodman, S., Grimsley, G., Layko, R., Melville, G., Mulder, M., Oberman, R., Panganiban, J., Peck, A., Seitz, L., Shea, S., Slevin, H., Youngerman, R., & Hobbs, L. (2020). geoBoundaries: A global database of political administrative boundaries. PLOS ONE, 15(4), e0231866. https://doi.org/10.1371/journal.pone.0231866

[36]. Salas-Olmedo, M. H., Moya-Gómez, B., García-Palomares, J. C., & Gutiérrez, J. (2018). Tourists' digital footprint in cities: Comparing big data sources. Tourism Management, 66, 13-25. https://doi.org/10.1016/j.tourman.2017.11.001

[37]. Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. ACM Transactions on Database Systems, 42(3), 19. https://doi.org/10.1145/3068335

[38]. Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(1), 72-101. https://doi.org/10.2307/1412159

[39]. Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. Scientific Reports, 7, 17615. https://doi.org/10.1038/s41598-017-18007-4

[40]. UNESCO World Heritage Centre (n.d.). Sudan: Properties inscribed on the World Heritage List. https://whc.unesco.org/en/statesparties/sd

[41]. UNESCO World Heritage Centre (n.d.). Archaeological Sites of the Island of Meroe. https://whc.unesco.org/en/list/1336/

[42]. UNESCO World Heritage Centre (n.d.). Gebel Barkal and the Sites of the Napatan Region. https://whc.unesco.org/en/list/1073/

[43]. UNESCO World Heritage Centre (n.d.). Sanganeb Marine National Park and Dungonab Bay-Mukkawar Island Marine National Park. https://whc.unesco.org/en/list/262/

[44]. van Zanten, B. T., Van Berkel, D. B., Meentemeyer, R. K., Smith, J. W., Tieskens, K. F., & Verburg, P. H. (2016). Continental-scale quantification of landscape values using social media data. Proceedings of the National Academy of Sciences, 113(46), 12974-12979. https://doi.org/10.1073/pnas.1614158113

[45]. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., & Bright, J. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17, 261-272. https://doi.org/10.1038/s41592-019-0686-2

[46]. Wikidata contributors (n.d.). Wikidata. https://www.wikidata.org/

[47]. Wikimedia Foundation (n.d.). Wikimedia Analytics API. https://doc.wikimedia.org/generated-data-platform/aqs/analytics-api/

[48]. Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. Scientific Reports, 3, 2976. https://doi.org/10.1038/srep02976

[49]. Zook, M., Barocas, S., boyd, d., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. PLOS Computational Biology, 13(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399