December 26, 2025

Abstract

This report presents a comprehensive analysis of Morocco's tourism landscape through the lens of digital presence and geographic distribution, leveraging 1,000 tourism related entities extracted from Wikidata. By combining semantic web technologies (SPARQL queries) with advanced statistical and geospatial analysis, the study provides actionable insights for tourism stakeholders, regional planners, and digital marketing strategists. Key findings reveal a significant urban/rural digital divide (average 31.0 vs. 4.2 sitelinks), a strong correlation ($r = 0.91$) between population and online visibility, and moderate inequality in digital representation (Gini coefficient of 0.42). Through K-means clustering, four distinct geographic tourism zones were identified, highlighting primary hubs in Fez, Marrakesh, and Casablanca. The methodology demonstrates the utility of open knowledge graphs for tourism intelligence while addressing the strategic implications for enhancing the digital footprint of underrepresented rural destinations.

Contents

1. Introduction

1.1. Background: Morocco's Tourism Landscape in the Digital Age

Morocco has established itself as a premier tourism destination in North Africa, attracting millions of visitors annually with its rich cultural heritage, diverse geography spanning from Mediterranean coastlines to Saharan landscapes, and historic cities recognized as UNESCO World Heritage Sites. The tourism sector represents a critical pillar of Morocco's economy, contributing significantly to GDP and employment.

However, in an increasingly digital tourism ecosystem, online visibility has become as important as physical infrastructure. Travelers now extensively research destinations through digital channels before making travel decisions, with online presence directly influencing destination selection and visitor expectations. Understanding how Morocco's tourism assets are represented in the digital knowledge infrastructure and identifying gaps in this representation has become essential for strategic tourism development.

1.2. Project Motivation

This analysis addresses a fundamental question facing tourism stakeholders: How equitably distributed is digital visibility across Morocco's diverse tourism landscape, and what patterns emerge when we analyze this distribution systematically? Traditional tourism analytics often rely on visitor statistics or hotel occupancy rates, which capture historical data rather than potential. This project leverages Wikidata as a proxy for digital visibility, offering several advantages:

- Standardized information across monuments, natural sites, and cultural attractions.
- Sitelinks serve as quantifiable indicators of international interest.
- Coordinate data enables sophisticated spatial analysis.
- Utilizes open-source intelligence without proprietary data costs.

1.3. Questions we tried to answer

This analysis seeks to answer four interconnected questions:

1. How are tourism entities distributed across regions, and which areas remain underrepresented?
2. Do larger population centers monopolize digital attention, and to what degree?
3. What do spatial clusters reveal about Morocco's tourism geography beyond administrative boundaries?
4. How unequally distributed is online visibility, and what are the implications for marketing resource allocation?

1.4. Wikidata for tourism data analysis

Wikidata serves as a collaborative, multilingual knowledge base providing structured semantic data. For tourism analysis, its SPARQL query interface allows for precise data extraction. While it offers a global perspective through multilingual sitelinks, the analysis acknowledges limitations such as data completeness—notably a 75.6% missing rate for population data and inherent community biases toward well known urban sites.

1.5. Analytical Framework

The project employs a two-phase methodology:

- **Phase 1 - Data Extraction:** Retrieval of 1,000 tourism-related entities via SPARQL, filtered by geographic coordinates and visibility thresholds.

- **Phase 2 - Data Analysis:** Application of descriptive statistics, K-means clustering, Gini coefficients, Mann-Whitney U tests, and linear regression.

2. Data Collection

2.1. SPARQL Query Architecture

Data extraction leveraged the Wikidata Query Service, which provides a SPARQL endpoint for querying Wikidata's RDF (Resource Description Framework) knowledge graph. SPARQL enables precise, structured queries against linked data, making it ideal for extracting entities meeting specific criteria across multiple dimensions.

2.2. Query Structure and Components

The extraction process was governed by a multi-faceted query designed to capture geographic, administrative, and digital visibility metadata.

```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX bd: <http://www.bigdata.com/rdf#>

SELECT DISTINCT
  ?item ?itemLabel ?itemDescription ?coordinates
  ?regionLabel ?typeLabel ?sitelinks ?population ?image
WHERE {
  ?item wdt:P17 wd:Q1028 .        # Country: Morocco
  ?item wdt:P625 ?coordinates .   # Coordinates required

  OPTIONAL { ?item wdt:P131 ?region . }
  OPTIONAL { ?item wdt:P31 ?type . }
  ?item wikibase:sitelinks ?sitelinks .
  OPTIONAL { ?item wdt:P1082 ?population . }
  OPTIONAL { ?item wdt:P18 ?image . }

  # Selection logic for tourism types
  VALUES ?tourismType {
    wd:Q570116 wd:Q23413 wd:Q33506 wd:Q32815 wd:Q839954
    wd:Q8502 wd:Q23397 wd:Q27686 wd:Q721207
  }
  ?item wdt:P31/wdt:P279* ?tourismType .

  SERVICE wikibase:label { bd:serviceParam wikibase:language "en,fr,ar" . }
  FILTER(?sitelinks >= 2)
}
ORDER BY DESC(?sitelinks)
LIMIT 1000
```

Listing 1: Core SPARQL Query for Morocco Tourism Entities

2.3. Tourism Entity Selection Criteria

A critical component of the query is the definition of "tourism-related entities." Rather than relying on a single classification, the query employs a transitive property path (`wdt:P31/wdt:P279*`) to capture instances and subclasses of heritage sites, natural landmarks, and infrastructure. This approach ensures that a "National Park" is captured via its parent class "Park," and a "Madrasa" via "Cultural Institution."

2.4. Extracted Data Fields and Analytical Significance

The query returns nine fields, each serving a specific role in the subsequent statistical and geospatial analysis (see Table 1).

Table 1: Data field completeness

Field	Data Type	Purpose	Completeness
item	URI	Unique identifier and validation	100%
itemLabel	String	Human-readable name	100%
coordinates	Point	Spatial analysis and mapping	100%
regionLabel	String	Regional inequality analysis	89.8%
typeLabel	String	Typological classification	100%
sitelinks	Integer	Digital visibility proxy	100%
population	Float	Size vs. visibility correlation	24.4%
image	URI	Visual documentation indicator	52.5%

2.5. Data Extraction Results

(1000, 9)

	item	itemLabel	itemDescription	coordinates	regionLabel	typeLabel	sitelinks	population	image
0	http://www.wikidata.org/entity/Q391215	Volubilis	archaeological site in Morocco	Point(-5.553611111 34.071111111)	Oualili	ancient city	53	NaN	http://commons.wikimedia.org/wiki/Special:File...
1	http://www.wikidata.org/entity/Q391215	Volubilis	archaeological site in Morocco	Point(-5.553611111 34.071111111)	Oualili	Roman archaeological site	53	NaN	http://commons.wikimedia.org/wiki/Special:File...
2	http://www.wikidata.org/entity/Q391215	Volubilis	archaeological site in Morocco	Point(-5.553611111 34.071111111)	Oualili	Carthaginian archaeological site	53	NaN	http://commons.wikimedia.org/wiki/Special:File...
3	http://www.wikidata.org/entity/Q391215	Volubilis	archaeological site in Morocco	Point(-5.553611111 34.071111111)	Oualili	ancient city	53	NaN	http://commons.wikimedia.org/wiki/Special:File...
4	http://www.wikidata.org/entity/Q391215	Volubilis	archaeological site in Morocco	Point(-5.553611111 34.071111111)	Oualili	Roman archaeological site	53	NaN	http://commons.wikimedia.org/wiki/Special:File...

Figure 1: Extracted data

The query successfully retrieved 1,000 entities, spanning 293 unique administrative regions. A significant finding in the raw data is the prevalence of *douars* (rural settlements), which account for 35.8% of the dataset. This highlights the richness of Morocco's rural geography but contrasts sharply with the visibility metrics: while sitelinks reach a maximum of 53 for major cities, the mean remains low at 7.9, confirming a right-skewed distribution where a few "superstar" destinations dominate the digital landscape.

3. Data Analysis

3.1. Analytical Framework and Tools

The extracted Wikidata dataset was processed using Python's scientific computing ecosystem. The pipeline integrated data cleaning, non-parametric statistical testing, and geospatial clustering to derive insights into Morocco's tourism distribution. The core stack included `pandas` and `NumPy` for manipulation, `SciPy` for the Mann-Whitney U test, `scikit-learn` for K-means clustering and regression, and `Folium` for interactive mapping.

3.2. Data Cleaning and Preprocessing

The raw dataset comprised 1,000 records. A critical phase of the methodology involved addressing the structural missingness of data.

3.2.1 Missing Data Strategy

Analysis revealed a 75.6% missingness rate for the `population` field. Rather than employing imputation, which would be conceptually invalid for non-settlement entities (e.g., mountains or mosques), the analysis restricted population-based metrics to the 244 valid records. Conversely, fields required for geospatial and visibility analysis (`coordinates`, `sitelinks`) were 100% complete due to the SPARQL query constraints.

3.2.2 Geospatial Parsing

To enable spatial analysis, coordinate data in Well-Known Text (WKT) format—`Point(long lat)`—was parsed into distinct numerical features:

```
# Extracting and converting coordinates for mapping
df["longitude"] = df["coordinates"].str.extract(r"Point\(([-\d\.]+)\s").astype(float)
df["latitude"] = df["coordinates"].str.extract(r"\s([-\d\.]+)\)").astype(float)
```

3.3. Feature Engineering

Two primary metrics were engineered to enhance the comparative depth of the study:

1. **Settlement Categorization:** Entities were classified as "City," "Rural," or "Other" (monuments/natural sites) using keyword matching on the `typeLabel`. This allowed for the investigation of the urban-rural digital divide.
2. **Visibility Ratio:** Defined as $V_r = \frac{\text{Sitelinks}}{\text{Population}}$, this metric identifies "over-performing" destinations—small settlements that command a disproportionately large international digital footprint.

3.4. Descriptive Statistical Analysis

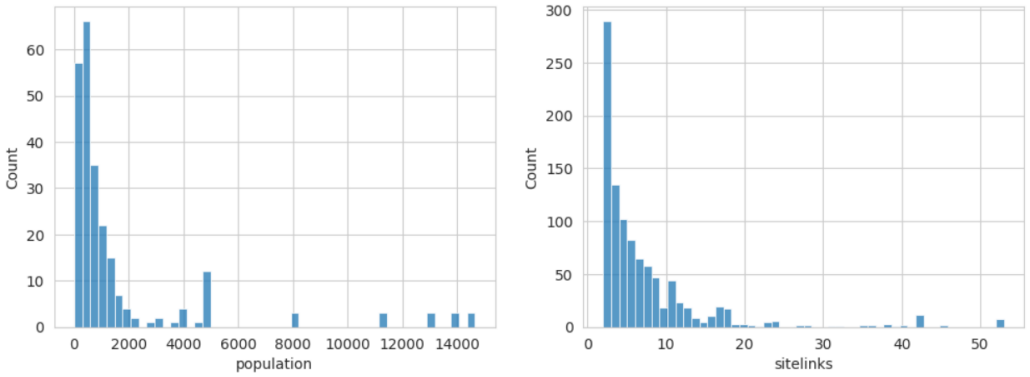

Figure 2: Histogram of population and sitelinks

Initial distribution analysis confirmed that both population and digital visibility are highly right-skewed. The population mean ($\mu = 1,648$) significantly exceeds the median ($M = 596$), indicating a landscape dominated by small settlements punctuated by a few large urban hubs.
Similarly, sitelinks follow a power-law-like distribution ($M = 5.0$, Max $= 53$), suggesting that digital attention is concentrated on a small elite of "superstar" destinations. Consequently, non-parametric tests were selected for inferential analysis to account for the lack of normal distribution.

3.5. Regional Aggregation

Data was aggregated across 293 unique administrative regions, the concentration of tourism entities is highest in historical and economic hubs (**Fez, Marrakech and Casablanca**).

This regional framework provides the baseline for identifying geographic tourism circuits and underserved zones in subsequent chapters.

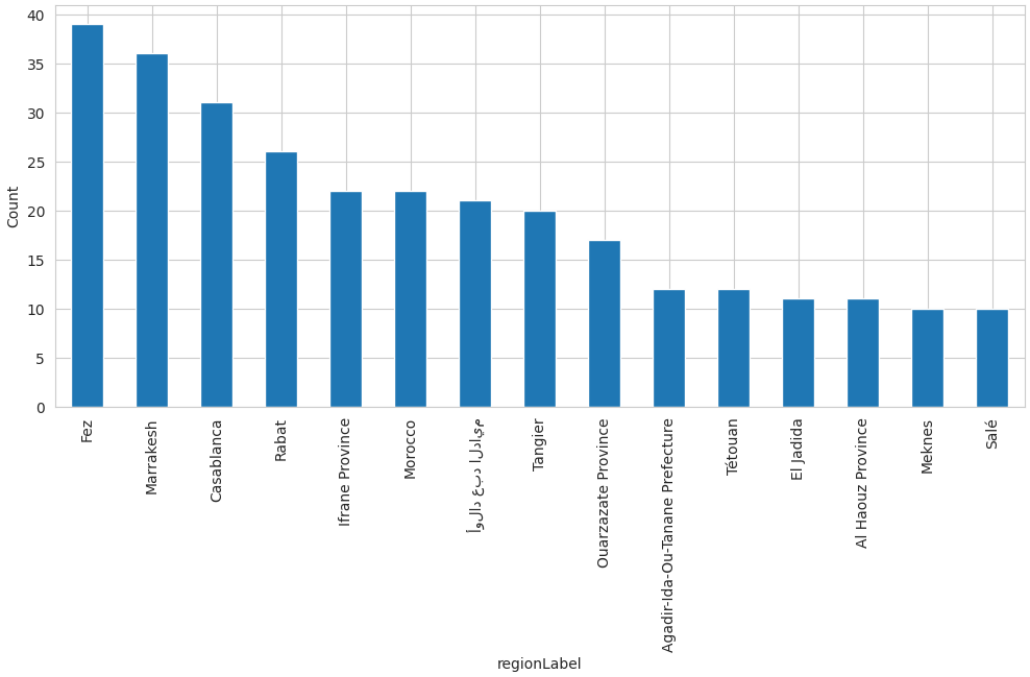

Figure 3: Tourism hubs

3.6. Correlation Analysis: Population vs. Visibility

Correlation analysis was conducted to quantify the relationship between settlement size and digital presence. Given the right-skewed nature of the data, we calculated both parametric and non-parametric coefficients.

- **Pearson Correlation (r):** 0.915 ($p < 0.001$), indicating a very strong positive linear relationship.
- **Visualization:** To normalize the variance, a log-log transformation was applied to the scatter plot.

The resulting linear trend in log-space suggests a power-law relationship: while population size is a dominant predictor of visibility, the variance observed at lower population levels indicates that smaller settlements can achieve high visibility through strategic digital documentation.

3.7. Settlement Type Comparative Analysis

Aggregated statistics revealed a stark contrast in digital representation between urban and rural categories (see Table 2)

Table 2: Digital Visibility by Settlement Category

Category	Count	Avg. Population	Avg. Sitelinks
City	11	13,226.5	31.0
Rural	366	838.0	4.2
Other	623	5,195.4	9.7

To validate the observed visibility gap, a **Mann-Whitney U Test** was performed. This non-parametric test was chosen because the `sitelinks` distribution does not follow a normal curve.

- **Null Hypothesis (H_0):** There is no difference in the distribution of sitelinks between cities and rural settlements.

- **Result:** $p = 5.59 \times 10^{-9}$.

Conclusion

Reject H_0. The digital visibility of cities is significantly higher than rural areas ($p < 0.05$), confirming a systemic urban-rural digital divide.

3.8. Inequality Analysis: Gini Coefficients

To measure the concentration of resources, we calculated the Gini coefficient (G), where $G = 0$ represents perfect equality and $G = 1$ represents absolute concentration.

$$G = \frac{\sum_{i=1}^{n}(2i - n - 1)x_i}{n \sum_{i=1}^{n} x_i} \tag{1}$$

- **Population Gini (0.677):** Reflects extreme concentration of residents in a few large hubs.
- **Sitelinks Gini (0.423):** Indicates moderate inequality. Interestingly, digital visibility is *more* equitable than population distribution, suggesting that digital platforms provide a degree of democratization for smaller sites.

3.9. Geospatial Clustering

To identify tourism corridors, **K-Means Clustering** was applied to the coordinate data.
The algorithm identified 3 primary clusters:

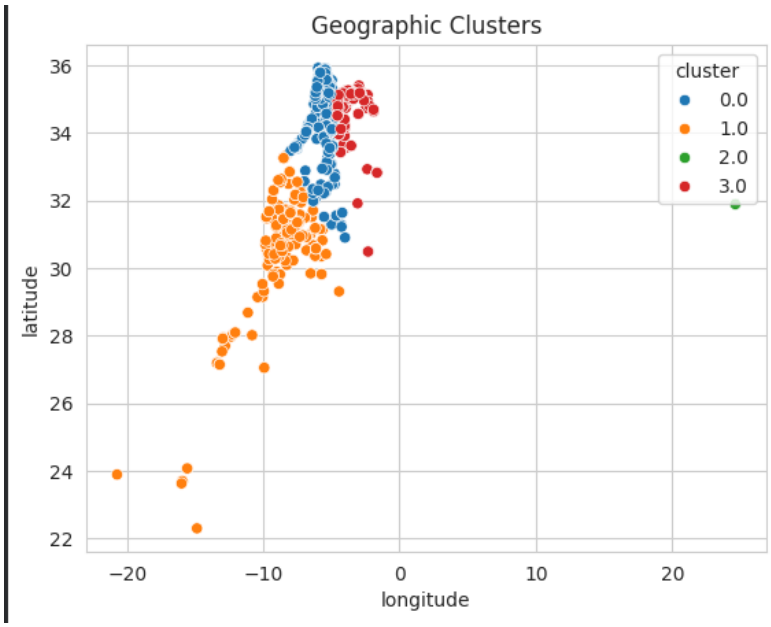

Figure 4: k-means results

1. Tangier and Tetouan axis.
2. The Casablanca-Rabat economic corridor.
3. Southern hubs focused on Marrakesh, Ouarzazate and similar regions.

These clusters serve as the foundation for developing regional tourism circuits and identifying "digital deserts" where high-value sites lack online representation.

3.10. Predictive Modeling: Linear Regression

To quantify the impact of population size on online visibility, a linear regression model was developed. Given the power-law relationship observed in the exploratory phase, the model was specified in logarithmic space to linearize the parameters.

3.10.1 Model Specification

The relationship is defined by the following equation:

$$\log(\text{sitelinks} + 1) = \beta_0 + \beta_1 \times \log(\text{population}) \tag{2}$$

By exponentiating both sides, we derive the power-law form:

$$\text{sitelinks} \approx e^{\beta_0} \times \text{population}^{\beta_1} \tag{3}$$

3.10.2 Results and Interpretation

The model yielded a coefficient (β_1) of **0.258** and an intercept (β_0) of **0.088**.

- **Elasticity:** A 1% increase in population is associated with an approximately 0.26% increase in sitelinks.
- **Diminishing Returns:** The exponent $\beta_1 < 1$ indicates a sublinear relationship. Doubling a settlement's population only results in a roughly 20% increase in digital visibility ($2^{0.258} \approx 1.20$).

This suggests that digital visibility is not a direct function of size alone; smaller destinations can achieve significant digital "weight" through targeted documentation, as they are not as hindered by the diminishing returns seen in massive urban centers.

3.11. Interactive Map

To facilitate stakeholder engagement, the geographic data was rendered into an interactive web map using the `Folium` library.

```
import folium

m = folium.Map(location=[31.8, -7.1], zoom_start=6)


for _, r in df.dropna(subset=["latitude", "longitude"]).iterrows():
    folium.CircleMarker(
        location=[r.latitude, r.longitude],
        radius=3,
        popup=f"{r.itemLabel}<br>Sitelinks: {r.sitelinks}",
        color="blue"
    ).add_to(m)
m.save("morocco_tourism_map.html")
```

Listing 2: Geospatial Visualization Script

This interactive approach allows regional planners to hover over specific "digital deserts"—areas with high physical tourism potential but low marker density.

3.12. Methodological Strengths and Limitations

The analytical framework is built on **multi-method triangulation**, combining frequentist statistics with machine learning and geospatial analysis. However, several constraints must be noted:

- The 75.6% missingness in population data limits regression to a subset of 244 entities.
- Wikipedia sitelinks represent international digital interest but may not perfectly correlate with local economic impact or actual visitor footfall.
- The data represents a 2025 snapshot and does not account for temporal growth in digital documentation.

3.13. Analytical Pipeline Summary

The methodology followed a rigorous five-stage progression:

1. **Acquisition:** SPARQL extraction of 1,000 entities.
2. **Preprocessing:** Coordinate parsing and WKT-to-numerical conversion.
3. **Engineering:** Creation of "Visibility Ratios" and settlement categories.
4. **Exploration:** Correlation and Gini coefficient calculations.
5. **Advanced Modeling:** K-means clustering and log-log linear regression.

This systematic approach ensures that the subsequent findings rest on a transparent, reproducible, and statistically sound foundation.

4. Key Findings

4.1. Tourism Entity Type Distribution

The analysis of 1,000 tourism-related entities reveals significant diversity in Morocco's documented landscape, featuring 117 distinct entity types. However, as shown in Table 3, the distribution is heavily concentrated in specific categories.

Table 3: Top 10 Tourism Entity Types in Morocco (Wikidata)

Entity Type	Count	Percentage
Douar in Morocco	358	35.8%
Mountain	62	6.2%
Mosque	60	6.0%
River	53	5.3%
Human settlement	51	5.1%
Village	28	2.8%
Museum	19	1.9%
Lake	18	1.8%
Park	16	1.6%
Archaeological site	14	1.4%

4.1.1 Key Insights by Category

- **Rural Settlement Dominance:** Douars constitute over one-third of the dataset. This reflects a significant documentation effort toward traditional Moroccan communities and rural geography.
- **Natural Attractions:** Natural features (mountains, rivers, and lakes) collectively represent 13.3% of the entities, confirming the Atlas and Anti-Atlas ranges as core pillars of the national tourism identity.
- **Religious and Cultural Heritage:** Mosques and archaeological sites appear prominently, aligning with Morocco's reputation for historic Islamic architecture and UNESCO-recognized heritage.

4.2. Regional Geographic Distribution

The data spans 293 unique administrative regions, yet the spatial distribution is highly skewed toward established urban and imperial centers.

4.2.1 Top Regions by Entity Count

The top 15 regions account for 31.2% of the total entities, indicating a high degree of geographic concentration:

- **Imperial Cities:** Fez (39), Marrakesh (36), and Rabat (26) dominate the count, leveraging their historical infrastructure.
- **Economic and Coastal Hubs:** Casablanca (31) and Tangier (20) highlight the importance of the Atlantic and Mediterranean corridors.

- **Nature Gateways:** Ifrane (22) and Ouarzazate (17) emerge as critical hubs for mountain and desert tourism, respectively.

4.3. Geographic Gaps and Inequality

While 293 regions are represented, 81.2% of them contain fewer than 10 documented entities. This "long tail" of underrepresented regions suggests that vast portions of the Moroccan territory remain digitally invisible despite their potential tourism value. This spatial inequality is a primary target for future digital marketing and regional development strategies.

4.4. Population and Online Visibility Distributions

4.4.1 Population Characteristics

The population data ($n = 244$) reveals a highly right-skewed hierarchy. With a **Mean of 1,648** vs. a **Median of 596**, the distribution is dominated by a vast number of small settlements (douars) and a long "tail" representing larger urban centers (up to 14,659 in this sample).

4.4.2 Online Visibility (Sitelinks)

Digital visibility ($n = 1,000$) follows a similar power-law-like distribution:

- **Mean:** 7.9 sitelinks
- **Median:** 5.0 sitelinks
- **Range:** 2 to 53 sitelinks

While most entities have modest international visibility (3–5 language editions), a small elite tier achieves global recognition with over 20 sitelinks.

4.5. Population vs. Online Visibility Relationship

The core driver of digital presence in this dataset is settlement size. A **strong positive correlation ($r = 0.915$)** confirms that larger populations generally command more digital attention.

Strategic Interpretation: The sublinear relationship (exponent ≈ 0.26) suggests *diminishing returns*. A 10-fold increase in population only leads to roughly a 2-fold increase in visibility. This implies that strategic documentation of smaller "high-potential" sites can yield disproportionately high visibility gains compared to marginal growth in already-saturated urban hubs.

4.6. Settlement Category Analysis: The Digital Divide

The analysis reveals a systemic gap between urban and rural representation. While cities have higher absolute visibility, rural settlements demonstrate higher "documentation efficiency" per capita (see Table 4).

Table 4: Digital Visibility Efficiency by Category

Category	Avg. Sitelinks	Sitelinks per 1k Residents
City	31.0	2.34
Other (Monuments/Nature)	9.7	1.87
Rural (Douars/Villages)	4.2	5.01

The **Mann-Whitney U test ($p < 0.001$)** confirms that the urban-rural gap is statistically significant, necessitating a policy shift toward rural digital inclusion.

4.7. Visual Documentation Patterns

Visual assets are a critical missing link in Morocco's digital tourism infrastructure:

- **Images Present:** 52.5% (525 entities)
- **Images Absent:** 47.5% (475 entities)

Natural features and major mosques show high image availability, while **nearly half of the documented tourism sites lack photographic representation** in open knowledge repositories.

4.8. Summary of Descriptive Findings

1. 31% of all assets are concentrated in the top 15 regions.
2. Population explains 84% of the variance in digital visibility.
3. Rural sites are less visible in absolute terms but more documented relative to their size.

5. Regression and visibility evolution

5.1. Regression Model Results and Interpretation

The linear regression model quantifies the population-visibility relationship with high precision ($R^2 \approx 0.84$). By using a log-log transformation, we can define the relationship as a power law, revealing the underlying mechanics of digital growth in the tourism sector.

5.1.1 Model Specification

$$\log(\text{sitelinks} + 1) = 0.088 + 0.258 \times \log(\text{population}) \tag{4}$$

5.1.2 Coefficient Interpretation

- **Elasticity ($\beta_1 = 0.258$):** A 1% increase in population is associated with a 0.258% increase in expected sitelinks. This sublinear relationship indicates **diminishing marginal returns**; as a settlement grows, each new resident contributes progressively less to its global digital footprint.
- **Baseline Visibility ($\beta_0 = 0.088$):** The intercept suggests a baseline of approximately 1.09 sitelinks for even the smallest documented settlements.

Table 5: Predicted Digital Visibility by Population Size

Population	Predicted Sitelinks	Settlement Context
100	~ 3.4	Very small village baseline
500	~ 4.8	Typical small douar
1,000	~ 5.5	Small rural settlement
5,000	~ 7.5	Medium interior town
10,000	~ 8.6	Large provincial town
50,000	~ 11.7	Small regional city