Travel Score – Open data for tourists

1. Introduction

Finding high-quality, tourism- and travel-related open datasets is a non-trivial task. While there are some high-quality datasets available,[1,2] they cannot be used freely (e.g. not commercially) and are therefore not open data as defined by the Open Definition.[3] This project aims to create such an open dataset, by combining open data from different sources. To prove the usability of the new dataset, the resulting data will be used to answer the following question:

Is North America more suitable for tourists than South America?

2. Used Data

2.1. Sources

The travel-score combined dataset is intended to serve as a basis for per-country travel score calculations. For that purpose, it has been composed out of four different data sources (see Table 1).

Data source	Purpose (in context of this project)	License
Inform Risk Index[4]	Risk profiles and indicators by country	CC-BY 4.0
International Comparison Program[5]	Price levels compared between countries	CC-BY 4.0
OpenStreetMap[6]	Map of tourism resources (natural, historic, etc.)	ODbL
Natural Earth[7]	Country boundaries & administrative details	PD

Table 1: Different data sources form the basis of the travel-score combined dataset.

2.2. Structure

For the scope of this project, the coverage was limited to North and South America. As such, the combined dataset contains a total of 35 records, one for each sovereign state within the Americas. For identification purposes, each record contains name, ISO 3166 code[8] and geometry data of the respective state, which have been taken from the Natural Earth dataset.

The rest of the data consists of numerical indicators: The top-level risk related scores have been copied from the Inform Risk dataset, while the data from the International Comparison Program has been used to include price levels for different product categories like food or transport. Several resource scores have been derived from the OpenStreetMap dataset (see Section 3.1 for more details).

Of all 35 records, 30 are complete, while the remaining 5 are lacking price levels. Depending on the respective data source, the numerical values have different domains. While risk indicators are constrained to the range 0 to 10, with 10 being the riskiest, price data is measured on a relative scale, with 100 being the global per-category average. Resource scores are measured on a similar scale but with 1 being the per-category average.

2.3. License

To comply with all source data licenses, different measures have been implemented. Most importantly, all datasets (except for Natural Earth) require attribution. To comply, attribution notices can be found in different places across this project and a feature to automatically add such notices to assets is planned. Another requirement becomes relevant when publishing resulting datasets: The share-alike terms of the ODbL mandate that any derived datasets must be published under the same or a compatible license.

3. Analysis

3.1. Calculation of resource scores

The OpenStreetMap is composed out of billions of different entities.[9] To handle this vast amount of data meaningfully, multiple computation steps were necessary. As such, the resource score calculation methodology significantly influences the analysis results and I decided to briefly discuss the steps here:

1. The target area is grouped into "GeoBins", each bin being a square with configurable side length (0.5° by default) on the equirectangular projection of earth.
2. For each bin, the number of different feature types (e.g. `tree`, `water`) per category (`natural` in this case) is counted. This way areas with increased resource diversity can be identified.
3. All bins get mapped to countries, then the counts are aggregated. The maximum became aggregation method of choice, because it performs well for countries with sparsely inhabited areas too.
4. The resource scores are calculated by normalizing the aggregates to an average value of 1.

3.2. Benchmarking

Before using the combined dataset to answer Question 1, some efforts should be taken to ensure that the data is accurate, i.e. it reflects the real world. For that purpose I benchmarked the combined dataset against the Travel & Tourism Development Index (TTDI).[2]

However, because the TTDI is licensed under the CC BY-NC-ND 4.0, I cannot publish the results without written approval. (Approval has been requested, as of yet an answer is outstanding.) Making the results available to individuals upon request should be permitted. Therefore, please contact me if you wish to receive further information regarding the benchmark results.

3.3. Calculation of travel scores

Using indicators from the combined dataset, the travel-score results can be computed. The basic idea is to calculate the overall score of a travel destination as the product of three independent factors, each one derived from the most relevant indicators in the combined dataset. The safety factor (Equation 2) for example is calculated from "hazard & exposure" and "lack of coping capacity" from the Inform Risk dataset, while leaving out "vulnerability" which is less relevant for tourists.

$$\text{total score} = \|\text{safety} \cdot \text{affordability} \cdot \text{attractiveness}\| \tag{1}$$

$$\text{safety} = \left\| \frac{1}{\text{hazard \& exposure} + \text{lack of coping capacity}} \right\| \tag{2}$$

$$\text{affordability} = \left\| \frac{1}{\text{actual individual consumption}} \right\| \tag{3}$$

$$\text{attractiveness} = \|\text{natural score} + \text{historic score} + \text{tourism score}\| \tag{4}$$

To prevent uneven scaling of factors, a specific normalization (denoted here as $\|.\|$) is used. This normalization is inspired by the min-max normalization, which maps values from the range $[\min, \max]$ to the range $[0, 1]$.[10] Different from the original, the minimum is always set to 0 and, assuming a normal distribution, the max is set to $2 \cdot \text{mean}(X)$. Outliers above 1 (which are rare in case of a normal distribution) can be clipped back into the range to improve consistency. Note that in Equation 2 and Equation 3 the inverse is taken, before the normalization is applied, to ensure that all scores share the same polarity (high values indicate pull, low values indicate push).

$$\|x_i\| = \min\left(\frac{x_i}{2 \cdot \text{mean}(X)}, 1\right) \tag{5}$$

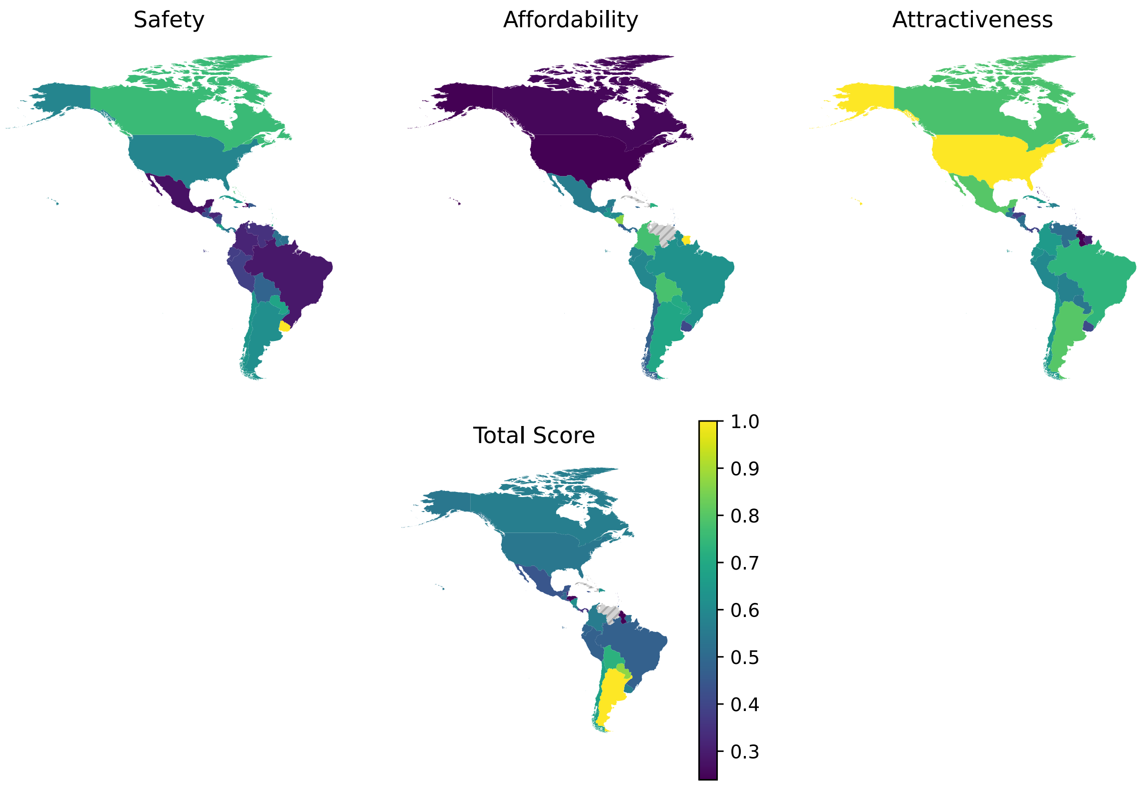

Figure 1: Visualizations of base factors and resulting travel score on a map of the Americas.

3.4. Discussion of results

The results, which have been visualized in Figure 1, will now be discussed in a geographical context. Please note that this context is only intended to give a quick overview and not to imply any causal relationships between geographic locations and scores.

Safety tends to be higher further away from the equator, with the safest mainland countries being Uruguay in the south and Canada in the north. Islands on the other hand are generally safer than most mainland countries, with Haiti and the Dominican Republic being the only exception.

South America is clearly the more affordable continent, considering that Canada and the USA are by far the most expensive countries, although some North American countries make it into top ten (Nicaragua, Dominican Republic and Guatemala). Attractiveness on the other hand tends to be a bit higher in North America, with the USA scoring first place, clearly before Argentina in the 2nd place and followed by more North American countries (Mexico and Canada).

So far this is not too far off from what might be expected based on common knowledge. However, the interesting part begins, once that all factors get combined to form the resulting travel score. Now Argentina emerges as the clear winner, directly followed by its neighbors Paraguay, Bolivia and Chile.

3.5. Interpretation

These results are especially interesting, because none of the three top countries score highest in any of the sub-categories, they simply do not score exceptionally low anywhere. This means that no factor managed to outweigh the others, which speaks for the efficacy of the employed methodology.

Additionally, among the three top countries, each one is best in its own sub-category (see Figure 2). This suggests, that in general there will always be the need for tradeoffs between different priorities, as no travel destination can be best in all aspects.

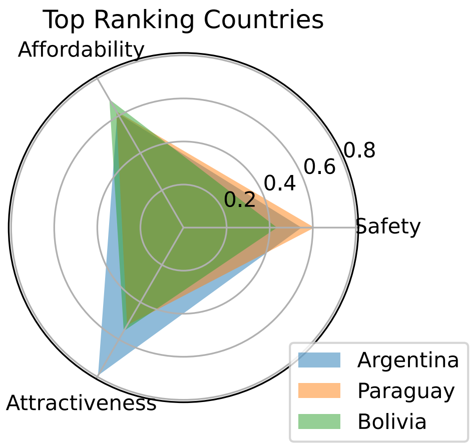

Figure 2: Visualizations of base factors for the three top ranking countries.

4. Conclusions

Based on my analysis, the answer to Question 1 is clearly *no*, with only South American countries ranking top three, far above the "classical" North American tourist destinations Canada (8th) and USA (11th). Many countries in South America are therefore definitely worth a visit.

4.1. Limitations in methodology

There are several shortcomings in the employed methodology, that could affect the validity of this answer. Most prominently, it is a bold endeavour to rank the attractiveness of a travel destination objectively, given that all tourists will have different personal preferences. For example, one could value safety far beyond affordability while others might argue contrarily. Likewise, it was not possible to account for all potential push and pull factors within the limited scope of this project.

4.2. Limitations in used data

Besides of the methodology, the underlying data might be inaccurate as well. As such, the data could be biased, especially in the context of OpenStreetMap. Here, volunteers contribute to a common data basis, which can lead to biases towards regions where many contributors live or visit.

Additionally, perfectly valid data in its original form could be misleading in a different context. The Inform Risk Index for example is constructed from the viewpoint of organizations, seeking an insight into the risk of humanitarian crises and disasters. This is obviously not the same as the risk for individual tourists, traveling through a foreign country.

Lastly, the aggregation of data at country level can be problematic. For example, the natural resource maximum might be far away from the historic resource maximum or the city at the touristic resource maximum is actually far more expensive, than the countries average costliness might suggest.

4.3. Outlook

Given these limitations, it becomes apparent that work on the methodology and the addition of data sources both have potential to increase the quality and credibility of the resulting dataset. A very promising approach would be the incorporation of risk and price data on a scale below country level. However, such datasets are usually less recent or do not exist yet at all.

Looking in the other direction, it would also be interesting to expand the existing analysis scope to global level, including other continents as well. While this would theoretically be possible using the present data sources, doing so might introduce new challenges and will consequently require additional tuning and verification work.

5. License Glossary

CC BY-NC-ND 4.0. Creative Commons Attribution NonCommercial NoDerivatives 4.0. https://creativecommons.org/licenses/by-nc-nd/4.0/ 2

CC-BY 4.0. Creative Commons Attribution 4.0. https://creativecommons.org/licenses/by/4.0/ 1

ODbL. Open Data Commons Open Database License. https://opendatacommons.org/licenses/odbl/ 1

PD. public domain. https://creativecommons.org/publicdomain/ 1

6. Bibliography

1. "UN Tourism - Tourism Statistics Database," <https://www.unwto.org/tourism-statistics/tourism-statistics-database>

2. "Travel & Tourism Development Index 2024," <https://www.weforum.org/publications/travel-tourism-development-index-2024/>

3. "The Open Definition," <https://opendefinition.org/>

4. "INFORM Risk," <https://drmkc.jrc.ec.europa.eu/inform-index/INFORM-Risk>

5. "International Comparison Program," <https://www.worldbank.org/en/programs/icp>

6. "OpenStreetMap," <https://www.openstreetmap.org/about>

7. "Natural Earth," <https://www.naturalearthdata.com/>

8. "The World Factbook - Country Data Codes," <https://www.cia.gov/the-world-factbook/references/country-data-codes/>

9. "OpenStreetMap - Database statistics," <https://taginfo.openstreetmap.org/sources/db>

10. "Codecademy - Normalization," <https://www.codecademy.com/article/normalization>