International Cartographic Association
Association Cartographique Internationale

From 75 Million Images to a Verified Dataset of Statistical Maps: A Multi-Stage Extraction Pipeline Combining Active Learning and Vision Language Models

Marta Solarz [a,*], Izabela Gołębiowska [a]

[a] *Department of Geoinformatics, Cartography and Remote Sensing, Faculty of Geography and Regional Studies, University of Warsaw, Warsaw, Poland – m.solarz2@uw.edu.pl, i.golebiowska@uw.edu.pl*

* Corresponding author

Keywords: statistical maps, benchmark dataset, MapPool, active learning, vision language models

Abstract:

Generative artificial intelligence offers a wide variety of support for cartography, including map design, map use, and evaluation (Kang and Wang, 2026). While developing machine learning architectures huge amounts of data are required for training, whereas for cartographic tools it often requires huge amount of maps of specific types. Our work-in-progress study aims at developing ML tools for statistical map quality assessment which requires a large, annotated dataset of verified statistical maps. Statistical maps are among the most common forms of geospatial data visualization, yet the cartographic quality of maps found online varies widely.

In this paper, we present a scalable, multi-stage pipeline for extracting selected types of statistical maps from MapPool (Schnürer, 2024a, b), a corpus of 75.9 million images classified as map-like from the CommonPool web-crawl dataset. MapPool contains a heterogeneous mix of actual maps of various types — topographic, navigational, thematic — as well as map-like graphics such as diagrams, infographics, or illustrations that resemble maps. From this diverse pool, our pipeline identifies and verifies a deliberately narrowed subset: statistical maps employing specific map types, intended as a benchmark for AI-based quality evaluation.

We define the scope of our study through three formal inclusion criteria. A candidate image must have sufficient resolution for visual analysis (C1), depict quantitative statistical data referenced to administrative or statistical enumeration units such as countries, regions, or census areas (C2), and employ choropleth mapping and/or circle-based proportional or graduated symbols mapping (C3). This deliberate scope restriction excludes other valid types of statistical maps (e.g., dot density maps, isolines, or cartograms) to ensure a well-defined and consistent evaluation framework. The applied criteria thus delineate our research scope rather than cover an exhaustive definition of statistical maps.

Our extraction pipeline comprises four stages (Fig. 1). To efficiently reduce the search space using readily available metadata, in Stage 1 (text-based pre-filtering), we score each of MapPool's 75.9 million images using only its textual metadata. We compiled 154 positive and 152 negative base phrases with cartographic relevance weights, translated them into 21 languages (covering 88% of the corpus), and matched the resulting 5,140 phrases against each record's text using an Aho-Corasick automaton. Images scoring below −1 are discarded, reducing the corpus to 54.9 million candidates (72.3% of the MapPool size). However, many images in MapPool have missing or uninformative text descriptions, which limits the discriminative power of keyword-based filtering alone — motivating the subsequent image-based classification stage. In Stage 2 (ML classification), we draw a stratified sample of 200,000 candidates and manually annotate 4,002 of them as statistical maps or not (Fig. 2A, B), skipping inaccessible URLs during annotation. Using MapPool's pre-computed CLIP ViT-L/14 embeddings (768 dimensions) as features, we compare several classifiers, with SVM-RBF achieving the best baseline F1-score of 0.82 on a fixed test set (N=801). One round of active learning with uncertainty sampling — annotating 600 additional images nearest to the decision boundary — improves the final model to F1=0.85, precision=0.76, recall=0.96, and PR-AUC=0.88. To scale the trained classifier beyond the annotated sample, in Stage 3, the final model is applied to all 54.9

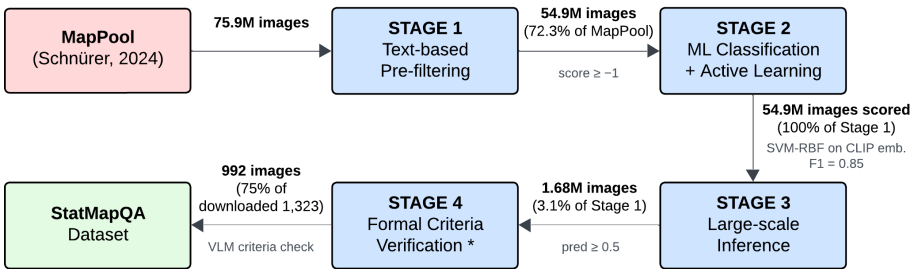

Figure 1. Multi-stage pipeline for extracting statistical maps from MapPool. Stage 4 (*) was applied to a pilot sample.

million pre-filtered candidates, yielding 1.68 million predicted statistical maps (3.1% positive rate). Because the embedding-based classifier cannot verify visual map content, a final criteria check on the actual images is required. In Stage 4 (formal criteria verification) approximately 35% of URLs were identified as no longer accessible, since MapPool originates from historical web snapshots.

Successfully downloaded images are evaluated against criteria C2 and C3 using a Vision Language Model with a structured prompt encoding our cartographic requirements. From 1,323 downloaded images, 992 (75%) pass all criteria, predominantly choropleth maps (91%, Fig. 2C), with proportional and graduated symbol maps comprising the remainder.

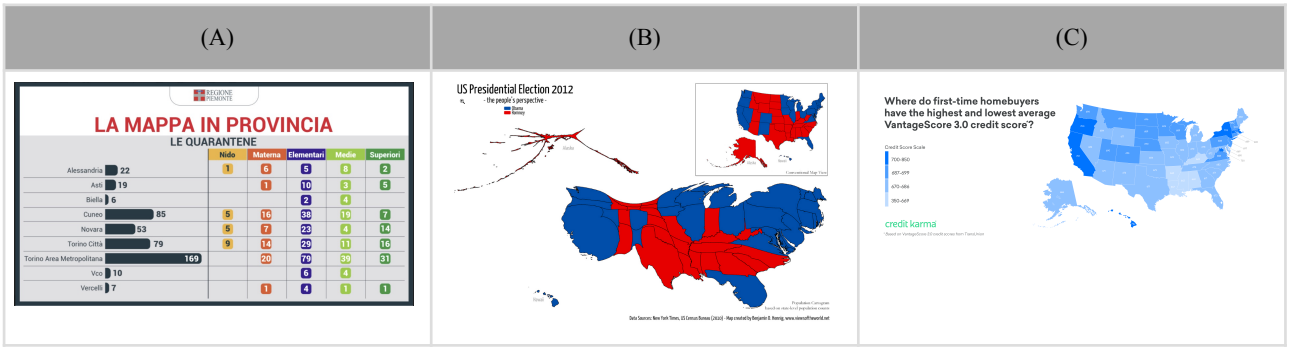

Figure 2. Example pipeline outcomes. (A) Statistical table — rejected (not a map, fails C2). (B) U.S. election map with categorical results — rejected (no quantitative data, fails C2). (C) Choropleth map of credit scores by U.S. state — verified (passes C1–C3).

This work contributes a replicable methodology for constructing domain-specific cartographic benchmarks from large web-sourced image corpora. The verified dataset — StatMapQA — will be publicly released with full annotations including criteria compliance, map type, and source metadata. The dataset serves as the foundation for the next phase of our research: developing an AI-based model for cartographic quality assessment.

References

Kang, Y., & Wang, C. (2026). Envisioning generative artificial intelligence in cartography: mapmaking, map use, and ethics. *International Journal of Cartography*, 1–27. https://doi.org/10.1080/23729333.2025.2582231

Schnürer, R. (2024a). MapPool – Bubbling up an extremely large corpus of maps for AI, 2024 ICA Workshop on AI, Geovisualization, and Analytical Reasoning – CartoVis24, Warsaw, Poland, *Abstr. Int. Cartogr. Assoc.*, 8, 22, https://doi.org/10.5194/ica-abs-8-22-2024.

Schnürer, R. (2024b). MapPool. HuggingFace. *https://huggingface.co/datasets/rschnuerer/MapPool.*