Towards Probabilistic Georeferencing of Geological Maps from PDF Reports

Marijan Soric
DI ENS, ENS, PSL University, CNRS, Inria
Paris, France
Supervised by: Cécile Gracianne, Ioana Manolescu and Pierre Senellart
marijan.soric@inria.fr

Abstract

Earth observations made by geologists are spatial and temporal; maps embedded in geologists' field reports are the primary record of where those observations were made. Recovering the geographic location of these maps is therefore a key data-mining step: it enables their export into a Geographical Information System (GIS) and makes the entire corpus of reports spatially queryable. Existing automated pipelines match the query against a fixed corpus of pre-georeferenced maps. This may limit coverage, and does not expose result uncertainty, even though every pipeline stage (OCR, NER, geocoding, image matching) is noisy. We propose an anchor-free probabilistic pipeline. A textual module extracts toponyms via NER on captions and map-specialized OCR on the image, then geocodes them into a spatial prior. A visual module scores candidate maps, generated on the fly from OpenStreetMap, against the query map using affine local feature matching between maps. We frame georeferencing as a black-box optimization problem and solve it with prior-guided Bayesian Optimization (πBO).

ACM Reference Format:
Marijan Soric. 2026. Towards Probabilistic Georeferencing of Geological Maps from PDF Reports. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (KDD).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Geologists document field trips in PDF reports that combine text (objectives, observations) with spatial and temporal data. A central piece of this data is the geological map, included as an image, which shows where the study took place. Extracting and georeferencing these maps automatically is essential for turning legacy reports into usable spatial data. Recent challenges such as GeolAug[1] (Inria and BRGM) and DARPA AI4CMA[2] reflect this growing interest.

Georeferencing a map is the assignment of geographic coordinates to a map image. This is a central problem in cartography and Geographic Information Systems (GIS) [8, 13]. Once solved, legacy maps can be overlaid on modern GIS layers and combined with other geospatial sources. Concretely, given a PDF report, the task is to detect the geological maps it contains and compute, for each map, a geographic transformation that aligns it to Earth's surface. We obtain this transformation by matching the query map against a candidate map with known coordinates: each match gives a Ground Control Points (GCPs), a pixel paired with its geographic coordinates. The map can then be placed in a GIS.

While early systems required users to manually supply control points[3], later work introduced semi-automatic pipelines for georeferencing [2–4, 7, 11, 12]. Recent fully-automated systems such as Uncharted[4] and DIGMAPPER [5] apply OCR to extract candidate place names, then match local image features against a reference corpus. DIGMAPPER, for instance, retrieves the closest match within a 10 km radius of a predicted center, from a corpus of pre-georeferenced US Geological Survey (USGS) maps. [15] follows a similar anchor-based approach for historical maps.

These methods share two limitations. First, they only work in regions already covered by a pre-georeferenced corpus, which excludes most of the world. Second, although their pipelines chain many components, each with its confidence score, uncertainty is not propagated end-to-end, making the final result hard to interpret or trust.

We propose a system that addresses both limitations in certain configurations. Rather than relying on a fixed corpus of anchor maps, we generate reference maps on the fly. We frame georeferencing as an optimization problem and solve it with Bayesian Optimization (πBO), guided by prior information (both textual and visual) extracted from the PDF report. Uncertainty is carried through the pipeline, so the final estimate comes with a confidence region.

2 Problem

Let $\mathcal{I}$ be a map image to be georeferenced, with known physical extent (the width and height of the area it depicts)[5]. Let $\mathcal{T}$ denote the text that may be associated to the map image: the figure caption, in-text references to the figure, and any labels detected on the map. We focus on the following sub-problem: given $(\mathcal{I}, \mathcal{T})$ and the physical extent of $\mathcal{I}$, recover the geographic center $x = (\lambda, \phi)$ of the area depicted. Solving this is sufficient for placing $\mathcal{I}$ in a GIS, and aligning the associated spatial data with other geographic layers.

[1] https://www.inria.fr/en/geolaug
[2] https://criticalminerals.darpa.mil/The-Competition
[3] https://docs.qgis.org/3.44/en/docs/training_manual/forestry/map_georeferencing.html
[4] https://github.com/DARPA-CRITICALMAAS/uncharted-ta1
[5] We neglect projection distortion.

KDD, Jeju, South Korea

ACM ISBN 978-1-4503-XXXX-X/2026/06
https://doi.org/XXXXXXX.XXXXXXX

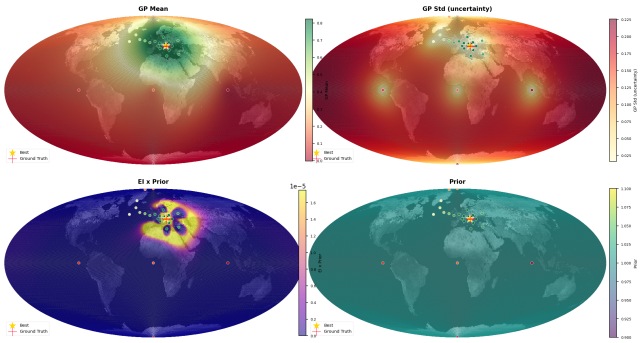

Figure 1: Visualization of πBO. (Top-left) GP mean: estimation of f. (Top-right) GP standard deviation: low near observations, high elsewhere. (Bottom-left) acquisition function: indicates where to sample x next. (Bottom-right) prior π (uniform prior).

3 Methodology

3.1 Preprocessing

We parse PDFs with Docling [1] to extract figures and surrounding text. For each figure we collect its caption and in-text references using simple regex patterns (*e.g., Figure X, Fig. X*). We classify each figure as either "geological map" or "other" using a Vision Language Model (VLM) (or, alternatively, a CNN/ViT classifier or zero-shot CLIP). The same VLM also can also be used to extract the map scale when it is explicitly stated in the figure, and otherwise returns an estimate.

3.2 Textual information

We extract candidate toponyms from two complementary sources. First, from the surrounding text: a Named Entity Recognition (NER) model tags location entities in the caption and in-text context, each with a confidence score. Second, from the map image: a map-specialized OCR system, mapKurator Spotter [10] detects text on the map, each detection with a confidence score.

Each candidate toponym is then geocoded using Nominatim[6] over OpenStreetMap[7], which returns a set of plausible Earth locations per name. At this stage we have multiple features: the OCR confidence, the NER score, the Nominatim importance score, and the Levenshtein distance between the candidate string and the matched name. These scores will weight the candidate locations during the optimization stage (subsection 3.5).

3.3 Visual information

Following prior work, given a candidate map of the same scale as the query $\mathcal{I}$, we run an local feature matching (LFM) model (LoFTR [14]) on the pair to produce pixel correspondences. RANSAC [6] then filters these correspondences to a geometrically consistent subset under an affine transformation. Although LFM models are trained on pairs from the same physical scene, they transfer reasonably well to map-to-map matching. The number of inliers, together with the LFM and RANSAC confidence scores, defines a similarity score between query and candidate. Figure 2 shows the inliers obtained on an example.

[6]https://nominatim.org/
[7]https://openstreetmap.org/

3.4 Bayesian Optimization over candidate centers

Let $\mathcal{X}$ denote the space of admissible candidate centers and let $f : \mathcal{X} \to [0, 1]$ be the similarity score defined in subsection 3.3. We assume the true map center satisfies $x^* \in \arg\max_{\mathcal{X}} f$. Unlike retrieval-based approaches that match $\mathcal{I}$ against a fixed corpus, we generate the candidate map at any $x \in \mathcal{X}$ on the fly. The function f is black-box (no gradients) and expensive: each evaluation requires rendering and running LFM + RANSAC. Naive grid or random search is expensive for large $\mathcal{X}$. We instead use Bayesian Optimization (BO): a surrogate model (Gaussian Process) is fit to past evaluations and an acquisition function selects the next candidate $x \in \mathcal{X}$ by trading off exploration and exploitation.

3.5 Prior knowledge

We use the textual evidence of subsection 3.3 as a prior over the location of the optimum. Define $\pi : \mathcal{X} \to [0, 1]$ as an approximation of $\pi(x) = \mathbb{P}[f(x) = \max_{\mathcal{X}} f]$. We model π as a mixture of 2-D Gaussians centered on the geocoded candidate locations $\mu_k = (\lambda_k, \phi_k)$:

$$\pi(x) = \sum_k w_k \mathcal{N}_2(x|\mu_k, \Sigma_k)$$

where the weights w_k are normalized aggregates of OCR, NER, Nominatim, and Levenshtein confidences scores associated. We then apply πBO [9], which reweights the BO acquisition function by π, concentrating search around textually-supported regions while preserving the ability to explore. Figure 1 illustrates the resulting BO trajectory on a test function.

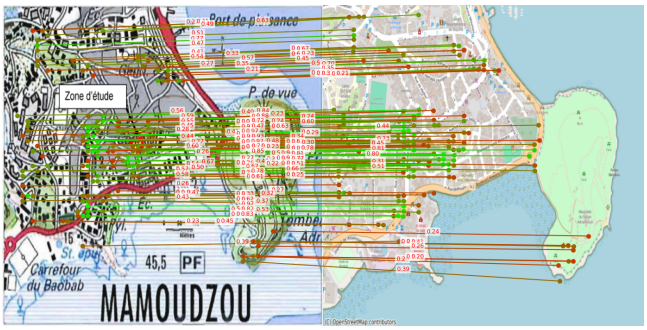

Figure 2: Results of LFM + RANSAC between (left) reference map from BRGM and (right) generated map from OSM

4 Discussion and Next Steps

We will evaluate on a benchmark of BRGM PDF reports with ground-truth georeferencing and on a synthetic OSM dataset, comparing against baseline with ablation (textual prior and BO search). Several extensions follow naturally: learning the mixture weights w_k, fine-tuning LFM across cartographic styles, and using VLM to extract richer visual and textual information priors.

References

[1] Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, Lokesh Mishra, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2024. Docling Technical Report. arXiv:2408.09869 [cs.CL] https://arxiv.org/abs/2408.09869

[2] James E Burt, Jeremy White, Gregory Allord, Kenneth M Then, and A-Xing Zhu. 2020. Automated and semi-automated map georeferencing. *Cartography and Geographic Information Science* 47, 1 (2020), 46–66.

[3] Ching-Chien Chen, Craig A Knoblock, and Cyrus Shahabi. 2008. Automatically and accurately conflating raster maps with orthoimagery. *GeoInformatica* 12, 3 (2008), 377–410.

[4] Yao-Yi Chiang, Craig A Knoblock, Cyrus Shahabi, and Ching-Chien Chen. 2009. Automatic and accurate extraction of road intersections from raster maps. *GeoInformatica* 13, 2 (2009), 121–157.

[5] Weiwei Duan, Yao-Yi Chiang, Theresa Chen, Michael P. Gerlek, Leeje Jang, Sofia Kirsanova, Craig A. Knoblock, Fandel Lin, Yijun Lin, Zekun Li, and Steven N. Minton. 2025. DIGMAPPER: A Modular System for Automated Geologic Map Digitization. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems* (The Graduate Hotel Minneapolis, Minneapolis, MN, USA) *(SIGSPATIAL '25)*. Association for Computing Machinery, New York, NY, USA, 717–728. doi:10.1145/3748636.3764602

[6] Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (June 1981), 381–395. doi:10.1145/358669.358692

[7] Mátyás Gede and Lola Varga. 2021. Automatic georeferencing of topographic map sheets using opencv and tesseract. In *Proceedings of the ICA*, Vol. 4. Copernicus Publications Göttingen, Germany, 38.

[8] Andreas Hackeloeer, Klaas Klasing, Jukka M Krisp, and Liqiu Meng. 2014. Georeferencing: a review of methods and applications. *Annals of GIS* 20, 1 (2014), 61–69.

[9] Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, and Luigi Nardi. 2022. πBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization. In *Tenth International Conference of Learning Representations, ICLR 2022*.

[10] Jina Kim, Zekun Li, Yijun Lin, Min Namgung, Leeje Jang, and Yao-Yi Chiang. 2023. The MapKurator System: A Complete Pipeline for Extracting and Linking Text from Historical Maps. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems* (, Hamburg, Germany,) *(SIGSPATIAL '23)*. Association for Computing Machinery, New York, NY, USA, Article 35, 4 pages. doi:10.1145/3589132.3625579

[11] Yan Li. 2006. Automated Georeferencing Based on Topological Point Pattern Matching. https://api.semanticscholar.org/CorpusID:11277540

[12] Jonas Luft. 2020. Automatic georeferencing of historical maps by geocoding. *Automatic vectorisation of historical maps* 13 (2020), 75.

[13] Kenzo Milleville, Steven Verstockt, and Nico Van de Weghe. 2022. Automatic georeferencing of topographic raster maps. *ISPRS International Journal of Geo-Information* 11, 7 (2022), 387.

[14] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. 2021. LoFTR: Detector-Free Local Feature Matching with Transformers. *CVPR* (2021).

[15] Beatrice Vaienti, Isabella Di Lenardo, and Frédéric Kaplan. 2026. Georeferencing historical maps using local feature matching and Delaunay consistency. *Cartography and Geographic Information Science* (2026), 1–23.