A Novel Approximate Computing Method for the Efficient Search in Satellite Remote Sensing Products

Ali Alsalama[1], Ahmed Kubba[1], Mohammad Alsmirat[1, 2], Isam Mashhour Al Jawarneh[1]
[1]Department of Computer Science, University of Sharjah, Sharjah, P. O. Box 27272, United Arab Emirates
[2]Department of Computer Science, Jordan University of Science and Technology, P. O. Box 22110, Jordan

Abstract— Remote sensing products include the generation of a class of geo-maps known as region-based geo-maps, such as choropleth and heatmaps. Comparing those maps is necessary for various real-time application scenarios and geo-maps time series analysis. A major challenge in this process is the vectorization of the raster images, transforming them into a compact data distribution format, in a way that reflects the color themes and densities of the source raster images which represent the geo-maps captured. Aggregation and grouping in such a process is indispensable, which is computationally expensive. To tackle this problem, in this paper, we showcase the design and prototyping of a novel efficient system GeoMapComp, for comparing a specific kind of remote sensing products efficiently, region-based aggregation geo-maps. We specifically compare geo-maps using proxies that are based on geohash encoding, where we apply geohash encoding to divide the geo-map area into equally-sized rectangles, then apply data distribution comparison metrics to compare those proxies, delineating then the differences between maps in a mathematically principled manner, incorporates the geographical characteristics of geo-maps, and is general-purpose and applicable to several kinds of region-based aggregate geo-maps. The paper further contributes by comparing several distance and point-based metrics such as Jenssen-Shannon, KL Divergence, and RMSE. Our results demonstrate the skills of our system in comparing region-based aggregate geo-maps remote sensing products effectively.

***Index Terms*—heat maps and choropleth, geospatial visualization, spatial data sampling, geohash, approximate query processing, Geospatial analysis; Data visualization; Geoscience, Jensen Shannon divergence, KL divergence, remote sensing, Map comparison**

I. Introduction

SATELLITE remote sensing (RS) has become an important tool in the field of geospatial analysis, as it offers the capability to acquire important data for monitoring environmental changes, managing natural resources, and exploring different domains and data distributions on color-coded geographic maps at a spatio-temporal level [8]. GeoTIFF images represent an important and common format for raster data, as they are a standard file format used in GIS applications and possess the ability to contain and visualize important information based on satellite observation [5]. Comparing different geospatial maps in the GeoTIFF format is an important part of geospatial data analysis and can be useful for many applications and processes such as urban planning [5, 6], which is why finding the best-performing and most consistent evaluation metrics for the comparison of GeoTIFF maps represents an important contribution to the field of research in geospatial analysis techniques.

The contributions of our paper consist of the following:

1) Comprehensive comparison of distance and point-based metrics such as RMSE, MAPE, Kullback–Leibler divergence, and Jenssen-Shannon Divergence for comparing a specific class of RS products (specifically region-based aggregate geo-maps) using a reference image and the generated samples.
2) Novel rasterization approach which preserves Geohash data in the rasterization process in the form of additional layers that store the Geohash coordinates and key (length), alongside pm2.5 values which are later color-coded for visualization.
3) Novel vectorization approach which recreates the original Geohash data from the multi-layered GeoTIFF images in order to compare the samples using their vector data. This method employs a novel conceptualization that we term as region-based geo-map proxy generator, which is responsible for generating relevant vector equivalents of the raster images, based basically on geo encoding approach known as geohash.

Corresponding author: Ali Alsalama (email: alialsalamaa@hotmail.com).

To compare the RS products, we first perform vectorization to perform the needed calculations for the evaluation metrics using each sample's pm2.5 values associated with their respective geohash row. The evaluated metrics consist of Root Mean-Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Jenssen-Shannon Divergence, and Kullback-Leibler (KL) Divergence. Essentially, this paper compares the performance of different distance and point-based statistical distribution metrics for comparing geo-maps (e.g., on the form of GeoTIFF images).

This paper is organized as follows: First, we cover the theoretical background necessary for understanding the paper's topics. Next, we present the design and prototyping of our innovative system, GeoMapComp, for region-based geo-map comparison and search. After that, we discuss the performance testing results obtained. Finally, we conclude the paper with some remarks and recommendations for future research.

II. Preliminaries

In this section, we provide an overview of the preliminaries and theoretical background on relevant topics essential for understanding the discussion on the design and implementation of our novel system presented in the paper.

A. Geo-visualization

Geo-visualization transforms georeferenced data into geographical maps through two main steps. First, geospatial data processing applies queries to georeferenced tuples [7], such as aggregating data into clusters using stateful operations like grouping and counting. This results in a dataset in vectorized geospatial format, which is then rasterized in the second step. Geospatial data visualization involves converting vector data into raster format and rendering it as geographical maps for user viewing [11].

There are three primary approaches for visualizing georeferenced data: point-based, line-based, and region-based . Point-based methods plot individual points like Points of Interest (POI), while line-based methods depict time-series trajectories of spatial data showing movement over time. Region-based approaches, which include generating choropleth maps, are particularly intensive in data processing. They tessellate geographic regions into grid cells and aggregate data using predefined regions such as neighborhoods or districts, assigning color density based on computed geo-statistics or aggregations.

Heatmaps are commonly used region-based geo-maps in smart cities. Despite variations, all region-based geo-maps require computationally intensive stateful data aggregation, particularly challenging in real-time data stream scenarios with fluctuating arrival rates. Georeferenced data is often parameterized during transmission to alleviate network load, representing spatial points as longitude/latitude coordinates, which can obscure their original geometrical context. Creating region-based maps involves converting these parameterized points back to their original form through geospatial aggregation, a computationally demanding process akin to geospatial joins in data stream environments . Effective geospatial data preprocessing is crucial for timely map generation. To mitigate delays, leveraging geospatial Approximate Query Processing (AQP) techniques like load shedding and sampling is recommended.

B. Map Rasterization

Map rasterization is the process of converting vector-based geographical data, which can consist of points, lines, and polygons, into raster images composed of a grid of pixels. This conversion allows for easier visualization and analysis of spatial data in systems that work with raster images, such as Geographic Information Systems (GIS) and image processing software. During rasterization, each geometric feature in the vector map is translated into one or more pixels on the raster image. This process involves determining which pixels correspond to the boundaries and interiors of the features. For example, a line feature might be rasterized by drawing pixels along its path, while a polygon feature would fill all the pixels within its boundary. Map rasterization is considered a fundamental process in GIS and RS applications, as it enables the conversion of complex vector data into a format that is suitable for analysis, visualization, and manipulation, allowing for the integration of various data sources and simplifying spatial analysis and modeling.

C. Map Vectorization

Map vectorization is the process of converting scanned or rasterized visual depictions of geographical objects, often termed instances, into a vectorized layout. This format allows for easier manipulation using Geographic Information System (GIS) software, which can lead to improved indexing, georeferencing, and spatial analysis. Vectorization represents a key component in extracting information from archival documents and plays a crucial role in enhancing the value of historical maps by enabling their usage in spatial and temporal analysis [2].

D. Geospatial Data Modeling, Reduction and Sampling

Handling large volumes of multidimensional data streams requires two key elements to ensure processing within the thresholds of Quality of Service (QoS) guarantees. These are geospatial data representations, followed by the imposition of access structures on these representations [19, 20]. Geospatial data representation models are divided in two main types: space-driven and data-driven. The space from which data is derived is usually called the embedding space. Space-driven models partition the embedding space (the geographical area from which data is derived) similarly to order-preserving hash functions, using methods such as quadtrees and grid files. In contrast, data-driven models partition the data items themselves, utilizing structures like R-trees and KD-trees. Space-driven grids can be either regularly sized equal grids or arbitrarily sized grids. This data representation is typically complemented by spatial data structures that facilitate quicker access to target data based on the spatial representation and distribution of the data [1].

To efficiently handle large volumes of geospatial data, it is essential to first represent the data using an appropriate multidimensional model, such as tree-based models. Following this, a spatial data structure, like B+-trees, is imposed on the model representation to enable fast data access. Modern GISs employ ”ordering” to streamline spatial data modeling and access. This technique reduces multidimensional data like geographic coordinates into single dimensions, often using Z-order curves. It organizes the embedding space (e.g., grids) and applies tree-based spatial access methods such as B+-trees on this ordered structure. This integration accelerates data retrieval, exemplified by geohash encoding in Z-order curve applications, simplifying spatial queries and enhancing efficiency.

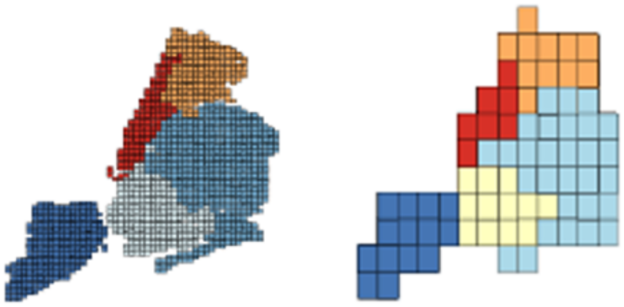
Geohash encoding serves as a textual representation for cells that depict the adjacent grid cell breakdown of an embedding space. Geospatial entities which share the same geohash prefixes belong to the same grid cell; the longer the shared prefix, the closer the spatial objects are in actual geometry. Geohash encoding is a useful tool for accelerating the handling of large volumes of geospatial data [4]. The length of the Geohash prefix is used as the key value for restoring the geohash from the raster file during map vectorization in this project, in addition to the center coordinates (longitude and latitude) of each individual geohash. Geohash geo-coding is considered an example of Geocoding, which is an approach that organizes

spatial data by transforming geographic coordinates into short strings. GIS use various formats for this purpose, taking into consideration factors such as convenience, performance, and legacy formats [9].

Geohash is a string representation denoting cells in an adjacent grid cell decomposition of an embedding space. Geospatial entities sharing identical geohash prefixes belong to the same grid cell; a longer shared prefix indicates closer proximity in actual geometry. Geohash encoding, along with other Z-order-based techniques like Google's S2 and Uber's H3, plays a vital role in expediting the handling of extensive geospatial data sets. In practical terms, geohash serves as a rapid and approximate filter for spatial proximity searches. Figure 1a shows the geohash covering for New York City (NYC) in the USA with a geohash precision of 6, while Figure 1b shows the geohash covering for the same city with a precision of 5. Precision is the number of characters in the string representation of a geohash value. For example, 'dr5rux' is the value of one of the grid cells covering NYC in Figure 1a, where 'dr5ru' represents a cell covering NYC at precision 5 as shown in Figure 1b. Geohash is used to condense geospatial data by converting geographic points into short strings, typically one to twelve characters long. Longer geohashes offer higher precision by representing smaller geographic areas. In AQP, sampling is a fundamental technique involving the selection of a subset from a population to estimate parameters like mean or count. The sampling design dictates how units or sites are chosen, aiming to create a sample that accurately mirrors the population's characteristics. While achieving perfect representation is impractical, the goal is to obtain a sample that realistically reflects the study variables with reasonable accuracy and confidence. One significant challenge in sampling is 'selection bias,' where certain parts of the population are systematically overlooked by the sampling method itself [6]. This bias can undermine the effectiveness of sampling designs, impacting the validity and generalizability of insights drawn from the sample.

There are two core sampling designs in the literature: simple random sampling (SRS) and stratified sampling (SS). In SRS, every unit in a population is assigned an equal selection probability. Labels are assigned to each unit, and labels are then selected randomly until the predetermined number of unique units, equal to the sample size, is captured. In SS, equal or non-equal portions are selected from each distinct group within the data, such as 50% males and 50% females from a student population in a school [6]. SS designs are preferred for their overarching benefits, as they are known to yield better estimations compared to random-based counterparts [6].

The rapid increase in voluminous geo-referenced data streams poses challenges for finding fixed solutions to complex spatial queries and real-time geo-map visualization. The multidimensional nature of geospatial data, along with its complex structures and varying data arrival rates, exacerbates these challenges. In geo-statistics and geo-visualization, there is a growing emphasis on Spatial AQP solutions that provide approximations with error bounds, which are highly valued in the literature [10]. To address these issues, spatial SS designs are recommended. This method efficiently selects samples from fast-arriving spatial data streams by dividing the population into strata based on specific criteria and independently sampling from each. This approach is effective for generating region-based approximate geo-maps from extensive geotagged data. SS is preferred over deterministic methods in scenarios where observing every object in a population, like monitoring migrating birds over large areas, is impractical. Overall, this method enables practical and efficient management of large-scale geospatial data analytics and visualization tasks, aligning with current research trends.



(a) With geohash precision 6 (b) With geohash precision 5

Fig. 1: Geohash tessellation for NYC city, USA

III. GeoMapCom: An Efficient Approximation-Based Method For Comparing Remote Sensing Products

This section describes the design and implementation of our novel system GeoMapCom, which we have designed for the efficient comparison of satellite RS products. We specifically focus on region-based aggregate geo-maps such as choropleth and heatmaps. We start by formalizing the problem of map comparison and search in the following subsection.

A. Problem Statement

Given a reference image of a map generated using RS (e.g., with the file format GeoTIFF) and a group of search images with the same file format. The scope of this paper focuses on the scenario in which the reference image and search images have the same resolution. The problem of map comparison can be conceptialized as follows: Find the image within the search images list which represents a map that is the most similar to the reference image (which represents the reference map).

B. System Design and Operation

To solve the problem described above, we have designed GeoMapCom which is composed of the necessary components which can synergistically work to find the closest image using minimal computational work.

Our novel system is composed of three main components as depicted in the context diagram of Figure 2. Those components are as follows: geospatial raster maps vectorizer, geospatial data modelling and representation module, stratified-like geospatial sampler, and map proxy creator. Our system operates as follows: The geospatial raster maps vectorizer is

responsible for transforming raster images that represent the maps (both the reference map and the search maps) into their equivalent vector representation. This results in data tables that contain, most importantly, each pixel in the map (represented as longitude/latitude pairs) and a number representing the density (from the color themes). Those data tables are then served to geo-encoder within the geospatial data modelling and representation module, which is responsible for generating geohash values representing each of the rows of the served table (from each longitude/latitude pair).

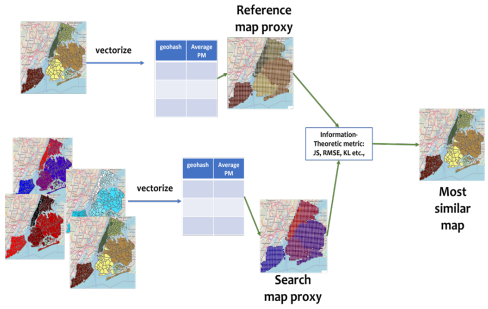


Fig. 2: GeoMapComp architecture

An attractive feature of the geohash is that it preserves spatial co-locality so that geographically nearby objects will have the same geohash values in the representation. Sampled data is then fed to the proxy creator. Map proxy is considered as a memory-efficient and compact representation of grouped geotagged vectorized data served as a matrix, where each row contains the statistic for each geohash, and the long/latitude pair of the geohash centroid point. This proxy is then passed to the comparison operator, which applies a statistical metric (e.g., Jenssen-Shannon, KL divergence, RMSE, etc.,) to find the image with the map that is the closest to the reference map among the search maps (specifically their raster image representations).

IV. Experimental Evaluation

In this section, we summarize test settings, the datasets, in addition to the baseline methods and evaluation metrics.

A. Experimental Setup

NYC AQ Dataset: We use the public Air Quality dataset from New York city in USA. It is a geo-referenced dataset (with longitude/latitude pairs of capturing locations) that consists of approximately 500K tuples. Air pollution is a significant environmental risk to urban communities, affecting all individuals, but with variations in pollutant emissions, exposure levels, and population susceptibility across different neighborhoods. Exposure to common air pollutants has been associated with respiratory and cardiovascular diseases, cancer, and premature death. The New York City Air Quality dataset contains metrics that offer insights over time and geographical areas within New York City, which aid in the comprehensive understanding of air quality and public health in the city. It includes data on various air pollutants such as particulate matter (PM2.5 and PM10) and is collected using low-cost

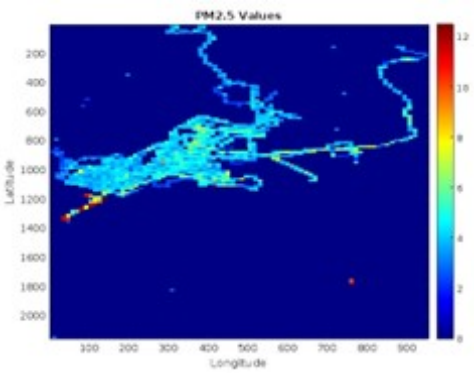


Fig. 3: Raster color map visualization sample

air quality sensors mounted on vehicles and circulating all through the NYC, spanning different time periods, such as hourly, daily, monthly, or yearly measurements.

Testing Procedure: Sample Generation, In order to perform experiments on the data, 50 GeoTIFF image samples were generated from the dataset for each sampling fraction (for a total of 500 generated images), each containing roughly the same Geohash codes but with different points randomly sampled from the dataset. SS was used to generate the images, with varying sampling percentages (from 10% to 90%) and, for one experiment, RS was exclusively used.

Map Rasterization and Vectorization as the NYC AQ dataset was in a vector format, it was necessary to rasterize each sample first to visualize the data on a color map. During rasterization, the average PM2.5 value was calculated for each geohash area, which is then used to visualize each geohash grid on the map with a color corresponding to the average PM2.5 intensity. In addition to the PM2.5 values, the geohash longitude, latitude, and string length (key) were added as three extra layers on top of the PM2.5 value, to facilitate the recreation of the geohash during the vectorization process. Vectorization of the samples is necessary to calculate the distance between a chosen reference image and the rest of the samples. Doing this requires subtracting the PM2.5 values from each image vector based on the corresponding geohash code.

The color map which can be observed in Figure 3 represents an example of a raster image sample generated from the preprocessed data based on the steps outlined above. As can be seen in the Figure 3, the minimum average PM2.5 value is 0, mapped to blue, and the maximum value is 12, mapped to the color red. The grids represented in this map are polygon squares generated from the geohash codes, each with a different color intensity based on the average PM2.5 value in the areas corresponding to the geohash-derived polygons.

Evaluation metrics: Four main metrics were used for calculating the distance between the reference image and the rest of the samples: RMSE, MAPE, Jenssen-Shannon Divergence, and the KL Divergence. In addition, the time spent on calculating each metric for the samples was also recorded using a built-in function of the IDE being used.

leftmargin=*

1) **Root Mean-Squared Error (RMSE):** Root Mean

Square Error (RMSE) is the Mean Squared Error (MSE) under the square root, to return the errors back to the original data's scale. It provides a more comprehensible measure of the average prediction error. A smaller RMSE indicates greater predictive accuracy, similarly to MSE, as it means that there is less distance between the two samples [7].
The RMSE formula is represented by the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

Where n represents the number of samples, Yi is the observed value, and yˆi is the matching estimated value [3].

2) **Mean Absolute Percentage Error (MAPE):** Mean Absolute Percentage Error (MAPE) is used to assess the accuracy of predictive models. It represents the average absolute percentage error, calculated by determining the absolute difference between the predicted value and the actual experimental value, dividing it by the actual experimental value, and then multiplying the result by 100.
The MAPE formula can be observed in the following equation:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \quad (2)$$

Where n represents the total number of observations, yi is the actual value at sample i, and yˆi is the predicted value for sample i.

3) **Jenssen-Shannon Divergence:** The Jensen-Shannon Divergence (JSD) is a measure of the similarity between two probability distributions. It quantifies the difference between two probability distributions by computing the average of the Kullback-Leibler Divergence (KLD) between each distribution and the average of the two distributions [11].
The following is the Jenssen-Shannon Divergence formula:
$D_{JS}(P||Q) = \frac{1}{2}\sum_i P(i)\log\left(\frac{P(i)}{M(i)}\right)$
$+ \frac{1}{2}\sum_i Q(i)\log\left(\frac{Q(i)}{M(i)}\right)$
Where P(i) and Q(i) represent the probabilities of the ith outcome under distributions P and Q respectively, while M(i) represents the average probability of the ith outcome.

4) **Kullback-Leibler (KL) Divergence:** Divergence (KLD), also known as relative entropy, is a measure of how one probability distribution diverges from a second, expected probability distribution. It quantifies the difference between two defined probability distributions [12].

TABLE I: Example of the NYC AQ dataset samples after pre-processing.

Latitude	Longitude	PM2.5	Geohash
40.847672	-73.869316	4.508813	dr72rh4
40.847668	-73.869316	5.462420	dr72rh4
40.847649	-73.869362	5.154881	dr72rh1
40.847649	-73.869362	4.508813	dr72rh1
40.847649	-73.869362	5.539503	dr72rh1

The below equation represents the KL Divergence formula:

$$D_{KL}(P||Q) = \sum_i P(i)\log\left(\frac{P(i)}{Q(i)}\right) \quad (3)$$

Where P(i) and Q(i), like in the Jenssen-Shannon Divergence, represent the probabilities of the ith outcome under the distributions P and Q.

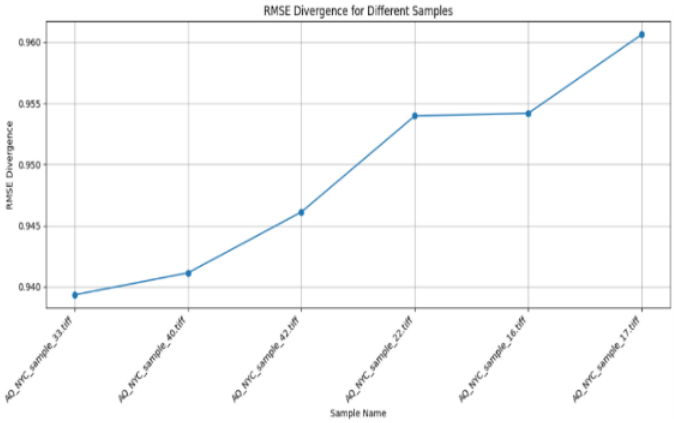


Fig. 4: Six closest images to the reference image using RMSE

Deployment: We deployed our experiments on a workstation on which the experiments and evaluations were conducted, running windows OS, and equipped with a 3050 RTX GPU with 4 GB of VRAM, an AMD Ryzen 7 processor, and 16 GB of memory RAM.

B. Experimental Results and Discussion

In this section we explore the usage of our GeoMapCom system in comparing RS products (specifically region-based geo-maps such as choropleth maps and heatmaps) generated from real big geo-referenced data. We depend on varying the sampling rate to generate search images. Then we use the error metrics (JS, KL, RMSE, etc.,) distance and point-based measurement to test the performance of the system and compare their skills in comparing images. We fix the geohash precision at 6, we obtain results shown in Figure 4 for the RMSE.

The graph observed in Figure 5 showcases the results of the ten experiments conducted using the RMSE, JSD, KLD, and MAPE metrics, with varying sampling fractions ranging from 10% to 90% and one SRS experiment. We applied a

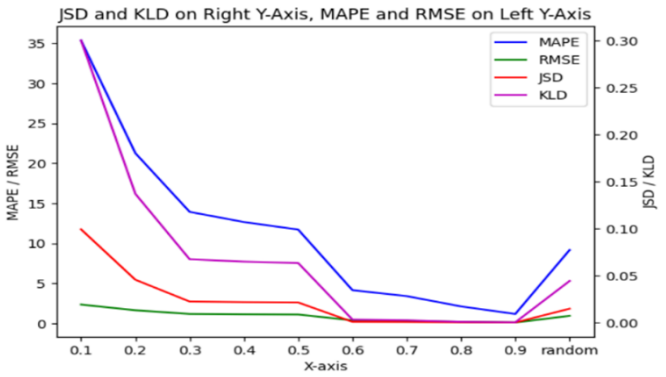


Fig. 5: Experimental results of the JSD, KLD, RMSE, and MAPE metrics

stratified-like sampling so that we generate images that are like the reference map with several degrees of accuracy. It is reasonable to assume that the higher the sampling fraction is, the closer the generated image would be to as compare with the reference image. Having said that, what we have observed in Figure 5 is that the RMSE decreases as we increase the sampling fraction (near linearly). A similar trend can be observed in the MAPE and KLD results, which decreases in value as the sampling fraction is increased with each experiment. However, in the case of MAPE and KLD, the decrease is much sharper and more notable compared to the RMSE which stays more consistent across the experiments.

As can be observed in the experimental results graphs for each metric, an almost consistent trend can be noticed across each sample fraction. As the sample fractions increase from 0.1 to 0.9, the distance metrics likewise consistently decrease, which can be explained by the fact that the higher the sampling fraction is, the more closely do the samples resemble the original GeoTIFF image, hence this fact is reflected in the distance measurements of each metric.

The highest recorded distance metrics are observed at the sampling fraction of 10%, whereas the lowest distance metrics occur at a sampling fraction of 90%. Additionally, no clear trend can be observed at the SRS fraction experiment, but it can be observed to give comparable results to the 30% to 50% sample fraction experiments across each recorded metric. A more elaborate investigation of the results is provided below.

As the sample fraction increases from 0.1 to 1.0, the Jensen-Shannon Divergence generally decreases. This indicates that the probability distributions of the samples become more like the reference distribution as the sample size increases. The highest value is at fraction 0.1 (approximately 0.099), showing the greatest dissimilarity. A sharp decrease is noticed from 0.1 to 0.2, after which the divergence values decrease gradually and more smoothly. The divergence is highest at the smallest sample fraction (0.1) with a value around 0.300. A notable decrease in divergence is observed up to fraction 0.5, beyond which the values level off, with minimal changes observed towards the largest fractions. The divergence for the random sample is notably lower than the smallest fractions but not as low as the highest fractions.

The trend across all metrics suggests that lower sample fractions are associated with higher errors or divergences. As the sample size increases, the metrics improve, reflecting better accuracy or similarity to the reference. This pattern underscores the importance of using adequate sample sizes in statistical analysis to minimize error and ensure reliable results. The SRS results suggest that while SRS can be effective, it might not consistently yield results as accurate as using the larger, more representative stratified sample fractions.

Figure 4 visualizes the six closest images to the reference image in one of the image searching experiments that utilize the RMSE metric. The Y-Axis represents the divergence (error) value of the image, whereas the X-Axis represents different map image samples. The closest sample to the reference image is the AQ NYC sample 33, with a divergence value of roughly 0.940. On the other hand, the most distant image is the AQ NYC sample 17 with a divergence value of 0.960.

V. Conclusion and Future Works

In this paper, we have shown the design and realization of a novel system that we term GeoMapCom for the efficient approximate comparison and fast search of a specific type of RS products, specifically those that resemble region-based aggregate geo-maps (such as choropleth and heatmaps). An overarching trait in GeoMapCom is that it reduces the RS products comparison process into a cheaper equivalent process by employing a proxy generator.

Storage, time, and computational constraints had to be taken into consideration during the experiments, which necessitated balancing the image resolution and hence quality with computational and time efficiency, including the size of the samples which can become considerably large with higher image resolutions making it difficult to store the experimental GeoTIFF samples.

For future directions, further research and work will be dedicated to developing a more efficient approach to preserving the Geohash data during rasterization, as the current approach in this paper suffers from a major limitation which is the fact that adding extra layers to the GeoTIFF samples in order to store the Geohash longitude, latitude, and key leads to a significant increase in the sample's size and hence causes additional storage, time, and computing constraints.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

References

[1] Isam Mashhour Al Jawarneh, Paolo Bellavista, Antonio Corradi, Luca Foschini, and Rebecca Montanari. Big spatial data management for the internet of things: A survey. *Journal of Network and Systems Management*, 28:990–1035, 2020.

[2] Yizi Chen, Joseph Chazalon, Edwin Carlinet, Minh Ôn Vũ Ngoc, Clément Mallet, and Julien Perret. Automatic vectorization of historical maps: A benchmark. *Plos one*, 19(2):e0298217, 2024.

[3] Dawei Duan, Shangbo Han, Zhongcheng Wang, Chunbo Pang, Longchao Yao, Weijie Liu, Jian Yang, Chenghang Zheng, and Xiang Gao. Multivariate state estimation-based condition monitoring of slurry circulating pumps for wet flue gas desulfurization of power plants. *Engineering Failure Analysis*, 159:108099, 2024.

[4] Isam Mashhour Al Jawarneh, Luca Foschini, and Antonio Corradi. Efficient generation of approximate region-based geo-maps from big geotagged data. In *2023 IEEE 28th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 93–98, 2023.

[5] Manolis Koubarakis and Despina-Athanasia Pantazi. *Legacy Geospatial Data Technologies*, page 31–52. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023.

[6] Sharon L Lohr. *Sampling: design and analysis*. Chapman and Hall/CRC, 2021.

[7] Mengyu Ma, Ye Wu, Xue Ouyang, Luo Chen, Jun Li, and Ning Jing. Hivision: Rapid visualization of large-scale spatial vector data. *Computers & Geosciences*, 147:104665, 2021.

[8] Emmanouil Oikonomou. Chapter 11 - remote sensing and geospatial analysis. In Nikolaos Stathopoulos, Andreas Tsatsaris, and Kleomenis Kalogeropoulos, editors, *Geoinformatics for Geosciences*, Earth Observation, pages 185–195. Elsevier, 2023.

[9] Pavel Petrov. Geocodes in geographic information systems. In *2023 International Conference Automatics and Informatics (ICAI)*, pages 225–229, 2023.

[10] Niklas Stoehr, Johannes Meyer, Volker Markl, Qiushi Bai, Taewoo Kim, De-Yu Chen, and Chen Li. Heatflip: Temporal-spatial sampling for progressive heat maps on social media data. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3723–3732. IEEE, 2018.

[11] Jia Yu. Src: geospatial visual analytics belongs to database systems: the babylon approach. *SIGSPATIAL Special*, 9(3):2–3, 2018.

[12] Yufeng Zhang, Jialu Pan, Li Ken Li, Wanwei Liu, Zhenbang Chen, Xinwang Liu, and J Wang. On the properties of kullback-leibler divergence between multivariate gaussian distributions. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 58152–58165. Curran Associates, Inc., 2023.