GLC-STATISTICS: A WEB-BASED SPATIAL STATISTICS SYSTEM FOR GLOBAL LAND COVER DATA

Ran Li[1]*, Wangzeng Liu[1], Yunlu Peng[1], Xiuli Zhu[1], Tingting Zhao[1], Linlin Che[1]

[1]National Geomatics Center of China, Beijing, China.- liran@ngcc.cn

Commission IV, WG IV/3

KEY WORDS: Global Land cover, Globeland30, spatial statistics, Web Service.

ABSTRACT:

Land cover data is one of the principle sources to understand changes of the earth, develop sustainable globe, conduct environment change studies, land source management, and many other societal benefit areas. Globeland30 is the first multi-temporal (for years 2000 and 2010) global land cover datasets at 30-meter resolution. Based on this dataset, an online geo-statistical system has been developed named "GlobeLand30 GSA". It helps users from different technical expertise backgrounds to visually explore and quantitatively analyse global land cover data. Different from other Web geoprocessing services and applications or desktop GIS software, this system has developed algorithms to correct multi-type errors in computing global land cover statistics, and applied optimal strategies of pre-organizing geography and auxiliary datasets in order to provide rapid response to multiple co-occurrence user-requests. This paper presents the system architecture and the developed data processing strategies, and also the related accuracy and efficiency assessment of GlobeLand30 GSA. The result shows that the system can provide accurate and efficient land cover statistical analysis for various size and shape of user requested areas, and is capable of handling as much as 200 concurrent user-requests with reasonable response time using a low-end desktop computer system environment.

1. INTRODUCTION

Land cover is one of the most important data layers for Digital Earth (Tateishi, et al. 2011), a critical earth observation priority for societal benefit (Zell et al., 2012), and a fundamental geospatial data set for implementing and informing sustainable development goals (Scot and Rajabifard, 2015). In the past ten years, a number of global land cover products at different spatial resolutions have been produced by international communities (Townshend et al., 2012; Chen et.al., 2015a) and widely used in sustainable development, environment change studies, land source management, and many other societal benefit areas (Mora et al.2014). Many of them data sets have been published on the internet and can be accessed by the users with the help of web-based functionalities, i.e., data browsing, downloading and visualization (Han et al., 2015; Brovelli et al. 2016). However, many users are not satisfied with these relatively simple services, but wish to have more on-line spatial analysis tools, such as spatial statistical analysis, validation, and information crawling. The development and provision of these sophisticated community-oriented services will enable land cover communities to promote resource sharing, perform value-added application, and enhance cross-board collaboration (Chen et al., 2017)

Spatial statistical analysis refers to the reduction of spatial patterns to a few clear and useful summaries by analysing the dependents of objects or events on their locations (Haining, 2014). As far as land cover is concerned, the acreage of individual cover types, magnitude of changes, spatial density or geographic concentration as well as degree of expansion or shrinkage are among the useful spatial statistics which can be derived. Several spatial statistics about global land cover and change have been reported, such as global urban expansion and built-up area (Angel et al., 2011; Chen et al. 2015), spatiotemporal pattern of global land surface water (Cao et al. 2014), distribution and diversity of global wetlands (WHIGHAM, 2009), global forest cover change (Hansen et al. 2013), Geographic distribution of global agricultural lands (Ramankutty et al. 1998). Most of these spatial statistical analyses were conducted using either GIS desktop toolkits or spatial analysis programming languages. For example, the geo-statistic module of ESRI ArcGIS integrating map algebra, geographical weighted regression and spatial auto-correlation is one of the commonly used toolkits. However, its utilisation for global analysis is quite complex and requires a careful design of optimal analysis strategies, such as appropriate processing regarding map re-projection and coordinates transform of vector data, or raster data re-sampling. Some programming languages (i.e., R language) were developed to support the geo-statistical analysis (Bivand et al., 2000). These programming languages can be freely accessed and widely used in specialized scientific communities However, significant programming efforts are required and may create barriers for end users. This is becoming even most difficult for ordinary users if they want to get land cover and change statistics of any particular area(s) in the world. It is therefore necessary to develop an online service system for facilitating the derivation of land cover and change statistics from the original land cover datasets.

On-line spatial statistics has been developed with the emergence of Web 2.0 and cloud computing technology, benefiting the Infrastructure as a Service (IaaS), Data Resource as a Service (DaaS) and Software as a Service (SaaS) (Yang et al., 2011). For example, the Joint Research Centre (JRC) of the European Commission developed a geo-statistical package which is used to interpolate or simulate users' observations point-data by using some classic geo-statistics algorithms like "ordinary kriging" automatically. This package is written with R language and has been implemented based on the Service Oriented Architecture (SOA) and can be freely accessed in a network environment (De Jesus et al., 2008). Since users need to upload their own geo-spatial data in XML file format via internet, it is inefficient to implement analysis on large volume point observations. The second example is CropScape (http://nassgeodata.gmu.edu/CropScape/), developed by National Agricultural Statistics Service (NASS) of the US Department of Agriculture (USDA) for on-line disseminating US

* Corresponding author

conterminous geospatial cropland data products (Han et al., 2012). It provides on-line statistical analysis function and enables the derivation of crop acreage and change statistics for any area of interest (Han et al. 2012). However, it is only for agriculture applications in the conterminous United States. The more recent example is the online platform for conducting spatial-statistical analyses of national census data across Australia (Petit et al., 2016). It has enabled users to identify access and conduct analyses of national census data via a single online platform. However, it was also limited to a national territory.

This paper presents the design and development of GLC-Statistics named "GSA (GlobalLand30 Statistics Application)", an on-line spatial statistical analysis system for the 30-m global land cover data set Globeland30. Globeland30 is a 30-m resolution GLC data product covering the entire earth land surface with ten land cover types and two baseline years (2000 and 2010) (www.globeland30.org) and was released for open access in Sept. 2014 (Chen et al. 2014). GLC-Statistics aims to allow users deriving accurate acreage statistics of land cover and change at any given geographic location or area through a user friendly interface with rapid response. Due to the global coverage and large data volume of GlobeLand30, on-line spatial statistical analysis at a global scale is much more complex than national or regional scale. Several critical issues have been identified and solved, including minimizing errors caused by global scale, efficient computation of large volume data. An optimal data management and processing strategy was proposed to implement GLC-statistic. The rest of paper is organized as the following: Section 2 discuses two critical issues related to GLC-Statistics, and section 3 presents the proposed methodology and algorithms. The performance of the GLC-Statistics was evaluated with real data and the results are given in Section 4. Section 5 discussed summaries this paper.

2. CRITICAL ISSUES AND THE STATE-OF-THE-ART

Conducting statistical analysis of 2 dimension (2D) spatial images, e.g., land cover maps, requests intensive computation effort because millions of pixels have to be processed to extract spatial information (e.g., detecting changes). This work load can be even larger when dealing with global dataset. Moreover, for different application purpose, researchers' region of interests (ROI) can vary from geographic locations, shape, size, and data acquired time, etc. It requires data processing algorithms to be robust that can be well adapted to various conditions to generate accurate statistical analysis results. Moreover, the implementation of the statistical analysis of land cover data as a user-friendly online service post extra requirements, e.g., quick response. Thus, these all, from result accuracy to computational working load, causes difficulties in developing an online statistic service facing. The details of multiple challenges in developing a suitable online global geo-statistics analysis service will be discussed in below.

2.1 Accurate estimation of geo-statistics at the global scale

User requested ROIs are normally determined based on the geography coordinates (a 3D system), and to calculate the corresponding geo-statistics, ROIs have to be projected in to a 2D coordinates. Thus, the key to estimate geo-statistics correctly is to solve the issue in approximating the 3D surface by projecting to a 2D surface. To be more specific, the projection residuals between 3D earth surface areas and 2D projected surface area (referred as projection error is below) are needed to be compensated in order to correctly calculate the size of requested areas at different geo-locations. Usually, the projection strategy is application-oriented. In other words, the projection method may vary from different application purposes and ROI attributes (e.g., regions with high/low latitude, large/small size of ROI regions). Hence, the conventional simple projection solution will not suitable for a system like Globeland 30 statistics that is supposed to handling user requests targeted any locations. An example is shown in Figure 1, one can see that area size have seriously changed after projecting the region in geography coordinate to UTM projection coordinate system, which is a type of isometric projection. Thus, the isometric projection is better to approximate the shape of the user-defined region of interest (ROI) in GlobleLand30 data, but it is not an optimal choice to minimize the projection error in the area-size calculation. In this case, using the isometric projection alone may not be enough to satisfy the accurate calculation required in geo-statistic calculation and evaluation. Hence, GSA has proposed and implemented a more sophisticated solution to keep the same accuracy standard for all type of user required ROIs. Our strategy combine the benefits of two projection strategies (Isometric projection and equal area projection), and adaptively determine processing parameters and iteratively compensate the existing projection errors. The algorithm details will be presented in Section 3.

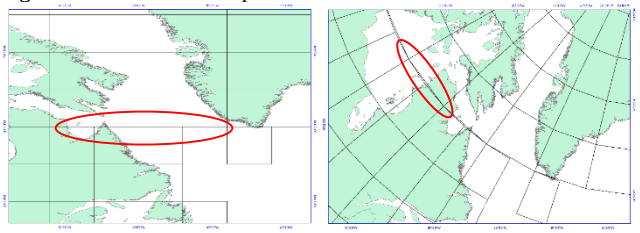


Figure 1 Demonstration of errors from projection schemes. (a) Geographic coordinate system; (b) UTM projection (a type of isometric projection). The red line circles the identical region in both geographic coordinate system and after UTM projection, to show errors in area-size after isometric projection.

2.2 Quick response of online web-service with large volume dataset

As an online web-service, user expect the statistics calculation of any defined ROI to be done in near-real time that is important for users' experience. However, different from the desktop geographical information system (GIS) software, the efficiency of online statistical analysis is restricted in many ways, the main impact factors including but not limited to (1) transmission time of data communication via internet (e.g., unstable/slow network); (2) analysis user requests and processing dataset (e.g., requesting processing large dataset); (3) computer and internet security policies, as the communication can be only made through internet explorer rather than direct access to computers; and (4) handling concurrent large number of user requests. The online service requires efficient computation algorithms to keep the result with sufficient accuracy, and quick response time between server and client (Fernandez et al., 2007). Moreover, to handle geo-spatial data that is commonly in the size of hundreds of Megabyte or larger, the global dataset is even larger, e.g., the compressed storage size of GlobeLand 30 land cover dataset is about 40 Gigabyte. The efficiency of the online system becomes even more critical, and development of efficient processing system are thus needed. An alternative solution is to improve the computation infrastructure, e.g., using cloud computing techniques. As the cloud system has less limitation on the storage room, computation cores and system memory. However, the cost of cloud service may limit the usage of scientific and non-profit applications. Thus, the online-service has to be constructed and maintained with affordable cost.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-4-2022-57-2022 | © Author(s) 2022. CC BY 4.0 License.

Additionally, in many examples of software development and system design cases, the result accuracy and computation efficiency are always incompatible. This is the same in calculating the geo-statistics of land cover maps. On the one hand the map projection error as well as other processing errors (e.g., raster data masking error that will be discussed in section 3.1 have be minimized in order to improve the accuracy and reliability of the statistical result. However, it would introduce more computational complexity and consume additional processing time. On the other hand, the online system need to response user requests quickly and consider the feasibility of transferring large volume data via internet. Thus, GSA implements a series of optimal calculation strategies of land cover data and also conduct accuracy lossless decomposition of user requests to balance the trade-off between computation accuracy and efficiency. Overall, our targeted online-statistics system should have not only compatible processing accuracy and efficiency, and it can be also well serves the scientific purposes and other non-profit applications. It could waive the software setup in user local computer, and relief the users from long dataset downloading time and the complex data processing steps.

3. AN ADAPTIVE DYNAMIC ITERATION (ADI) METHOD

The overall data processing strategy of the GSA is demonstrated in Figure 2. On the basis of the GlobeLand30 dataset and other auxiliary GIS datasets, the primary system design philosophy is to (1) minimize errors in the calculation of statistics parameters; (2) processes large volume of datasets; and (3) quickly respond to user requests. To maintain computation and response efficiency, the GSA will dynamically decompose the user requests (Section 3.1). To minimize the error in the data processing, the system implements an algorithm that iteratively calibrates intermediate data before the computation of final statistics (Section 3.2). Specifically, the core data processing steps developed and implemented in GSA includes schemes in source data organization, dataset partition and iterative method for result calibration, all of which will be discussed in details in the rest of this section.

3.1 Adaptive Dynamic scheme for task decomposition

As mentioned above, the efficiency is a key issue to construct a successful online system, to achieve this goal, an adaptive dynamic scheme has been developed and implemented for GSA. The scheme can (1) decompose the user requests based on the source data organization and the service-oriented computing (SOC) techniques, and (2) adaptively partition the dataset to further improve the computation efficiency. The rest of the sub-section will discuss these two parts in details.

3.1.1 Dynamic user request decomposition

First of all, source data organization (including GlobeLand 30 land cover data and GIS auxiliary datasets) together with SOC techniques have been used to decompose user requests, which have played an important role in reducing the data processing complexity and improving the result accuracy. Briefly speaking, all the datasets have been carefully organized and some statistical results have been pre-calculated and stored as well that is considered as a part of the static dataset. Afterwards, any coming user request will be dynamically divided into a combination of the parts need real time computation and the others that can be grabbed from static dataset. Thus, this pre-organization scheme and request decomposition service also build the first system logic unit that responses to the incoming user requests. The following discussion will provide detailed information on these two aspects.

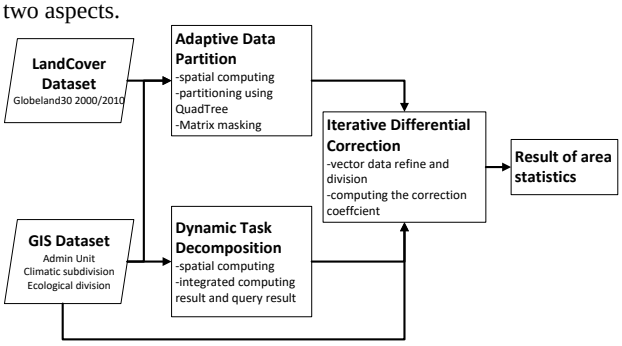


Figure 2. An overview of the data processing and management strategy. The rectangular box in the middle outlines the main optimized algorithms.

(1) source data organization

The GlobeLand30 data is stored in the non-destructive GeoTIFF compression format with 256 indexed 8-bit colour pattern, and consists of 5 parts, namely classification result files, coordinate information files, map setting file of classification image, metadata file and illustrative file. The data is based on the UTM projection with 6-degree zoning in WGS84 datum. According to areas with different latitude situations, two methods are adopted to organize the data tiles. For the area between 60 °N and 60 °S, the data tile is implemented in the size of 5 ° (latitude) × 6 ° (longitude). For the area between 60 ° and 80 ° of south-hemisphere and north-hemisphere, the data tile is saved in the size of 5 ° (latitude) × 12 ° (longitude) and the projection is conducted according to the central meridian of 6 °zone with an odd number. There is no data for the region above 80 ° latitude. Overall, there are 853 data tiles in total. Figure 3a shows the grid map of globeland30 dataset that bring the idea of total data tiles covering the whole globe.

Additionally, GSA has pre-calculate the area size of every land cover type for every 0.1 °×0.1 °latitude-longitude grid. The individual unit grid and its statistics result is connected through a spatial index. Thus, for every unit grid, the data management module links together the corresponding pre-calculated statistics data with land cover data and other pre-loaded geo-spatial data. This pre-calculated data, considering part of static dataset, is used to reduce the system processing efficiency. We have performed the same pre-calculation for 2000 and 2010 Globeland30 land cover data, all of them are stored in the server database that can be easily and quickly loaded when needed to initiate the dynamic service decomposition that will be discussed in below.

(2)Dynamic decomposition of user requests with SOC techniques

When the system receives a user request, the initial request can be decomposed into multiple-tasks that are sent to background processing modules. The system will determine the type of user requests: (i) if the requested statistics data already exists in the background database, e.g., statistics data for a whole country; (ii) if the calculation can be done through spatial relations between user's ROI and any pre-calculated dataset; and (iii) if real-time calculation work load can be reduced by cooperating the pre-calculated statistics data. Such decision making step is a dynamic process. Depending on various type of user requests, the GSA system will perform different computation algorithm.

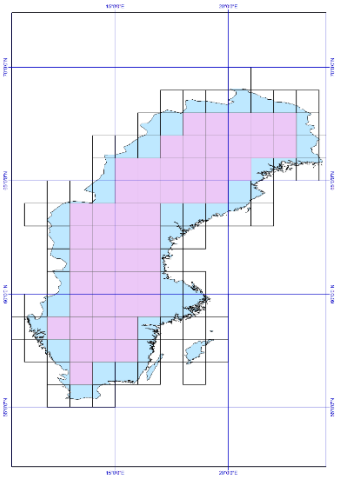
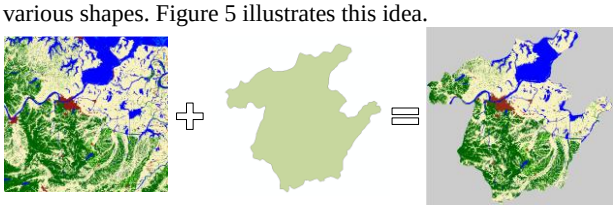


Figure 3. An example demonstrate the dynamic request decomposition.

For the first two types, the system will compare geo-coordinates of the user requested ROI region to the spatial-index system, then simply grab the result from the existing stored static dataset or compute the result through simple calculation (e.g., adding up). For the third case, the system will send the data to the next sophisticate processing step, namely dynamic matrix decomposition for zonal statistics, which will be described in the coming sub-sections. An example is given in Figure 4, it shows the decision making processes of GSA in decomposing user requests. Every individual request with irregular shape (e.g, Blue region in Figure 3) can be divided into two parts: combination of a stack of grids and the result of irregular areas for further calculation.

3.1.2 Adaptive Data Partition

After the determination of users request that needs real-time calculation, the ROI subset at land cover map will be determined and more important the region outside ROI region will be masked. The GSA system will first plot the ROI region at land cover dataset that equivalent the original user request, which can be done by a series of map projections and coordinate transformations from the spatial reference of the user requested ROIs to the matrix index of the stored land cover data. Afterwards, the GSA system can mask the subset in order to exclude regions outside the user requested area. Note that the original as well as the re-projected user requested ROI can be in various shapes. Figure 5 illustrates this idea.

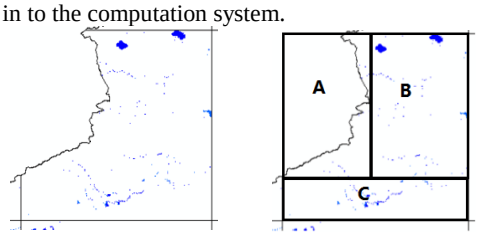


a. Globeland30 b. re-projected user ROI c. masked GLC30

Figure 4. Visualization of land cover data masking. Areas at Globeland 30 land cover maps outside user requested area will be discarded.

The key challenge here is to operate the mask and land cover dataset with the minimum computation load and best computation efficiency. As shown in Figure 4, normally a separate matrix in the same size as the land cover data matrix is used as the mask, which needs extra memory to build and sometimes can be large. For example, for a map sheet in mid-latitude region that is in size of 5 ° latitude ×6 ° longitude, the dimension of this matrix is about 17000 ×19000 pixels. Assuming there is 8-bit quantization for each pixel, it is needs about 700 megabits memory space. This requires a large amount of computer memory to be allocated to perform the matrix mask operation. Thus, it is a great burden for online land cover data statistics service, especially when multi-user concurrent requests and multi-tasking processes present.

We have developed an Adaptive Data Partition scheme to overcome this issue. Figure 5 illustrates the concept of this solution. First of all, the module establishes the mapping relationship between the requested ROI and the raster matrix of GlobeLand30. Then, the matrix is partitioned iteratively in a quadtree fashion, where the matrix is subdivided recursively into smaller regions, until the geographic boundary can be covered by a single matrix block with the minimized size. In Figure 6, the masked area is the block-matrix A, while the rest of the matrix can be calculated directly without additional masking operations. In this way, the mask matrix has been decomposed and minimized so as to avoid the step of allocating large memory, thus improving the computation efficiency. Overall, the key purpose of this algorithm to dynamically decompose the mask region, so that the system can avoid loading a very large mask matrix in to the computation system.



(a) (b)

Figure 5 Demonstration of Dynamic Matrix decomposition for zonal statistics. (a) is the area before the decomposition and (b) shows that after decomposition (A, B and C marks the different titles).

3.2 Iterative algorithm with differential corrections for result calibration

After the masked region is determined dynamically as discussed above, the system needs to count the number of un-masked pixel for every class. Again, users expect to obtain the result that represents the curved earth, while the stored geo-data are projected in 2D map. To connect the request and database, the system needs to apply an appropriate map projection scheme. The classic projection methods can be categorized into three groups, isometric projection, equidistant projection, and equal area projection. They can minimize errors of either angle, distance or area-size when projecting a small region. However, it is difficult to obtain an accurate result, if dealing with large areas (e.g., global land cover statistics analysis), both in size and shape with a simple and single project. Our solution in GlobeLand30 includes corrections of the projection deformation and the area distortion.

1) Correction for projective deformation of polygon

To outline the ROI region in land cover data and avoid unpredictable error cause by raster data resampling. It is a necessary step that converts the requested ROI in user defined spatial reference to the Globeland30 data spatial reference. In order to ensure the shape-invariant after transferring ROI to the spatial reference frame of the land cover data, isometric projection has been implemented. In this transformation and also for other projection methods, the errors introduced by discrete points need to be considered. As we know, GIS vector data are saved as discrete points, whose coordinates were sorted as serialized data, and are used to approximate the continuous polygons in projection transformation. An example is shown in

Figure 6, it demonstrates the errors raised projecting polygons that leads to overlaps or missing areas. This issue is particularly significant in the high latitude region or when requesting large ROI. In order to solve such problems, we adopt the idea of differential geometry of curves (Abbena et al., 2006), and implement it in GlobeLand30 geo-statistics system. Basically, we add multiple points to a line features in vector data and then implement the projection. As shown in Figure 7, after correction, the projected features can better approximate the real shape of the line.

2) Correction for area distortion

As mentioned above, when the requested region in land cover data is determined, the area-size of every land cover class can be calculated by counting un-masked pixels. This calculation is processed under the UTM projection system that is spatial reference of the Globeland30 land cover data. As the UTM projection is considered as an isometric projection, while the equal-area projection is supposed to be better in area-size computation. Thus, errors from counting area-size of a pixel in UTM projection needs to be corrected.

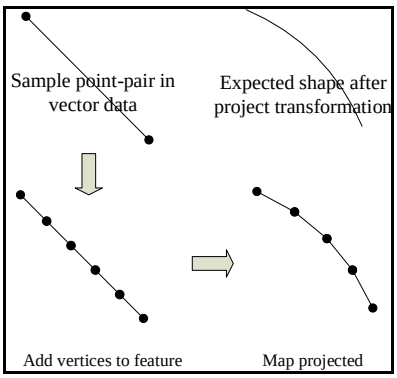


Figure 6. Demonstration of projective deformation errors and the correction algorithm with differential geometry theory.

Our approach for area distortion correction first divides the requested ROI into several regular grid cells, calculate each grid cell area in both isometric projection and equal-area projection, and determine the compensation factor for every grid cell that varies at different latitude locations. Then, the initial result of the ROI area-size is summarized from the overall size of pixels in UTM projection and the summation of the corresponding compensation factors in requested ROI. Afterwards, we can compare this initial result with the overall area-size of the ROI that is transformed in one-piece via equal-area projection. The residual in between the two is calculated. If the residual is smaller than a pre-defined threshold value, which is 0.3% in this system, the result is considered as acceptable. Otherwise, we further divide the grid cell in isometric projection into smaller grid and repeat the calculation until the residual is smaller than a threshold value. This iterative correction scheme is shown in flowchart in Figure 7.

3.3 Implementation of ADI online-statistics analysis

Overall, the main benefit of the carefully designed GSA statistic system is that one can optimize the calculation strategies and improve the computation efficiency despite the area size, or geography location, or irregularity of ROI shape of different user's request and maintain the same result accuracy. GSA decomposes individual user request into a dynamic part, the computational efficiency can be largely improved and the user experience can be enhanced. The dynamic matrix decomposing step help system minimize the size of mask matrix that can reduce the computation load and improve the real-time processing efficiency. The iterative algorithms with differential correction for statistic result takes the projection errors into considerations, and can guarantee the accuracy of the final result. Thus, GSA can process the various user requests, despite of their ROI size and geography location, and achieve the same accuracy level.

Other than the improvement in statistical data analysing algorithms, the GSA online-statistics system has also integrated the common web-map functions, such as layer display control and map exploring. So that user can freely visual analyse the GlobeLand30 land cover maps and other pre-load geo-spatial maps. The pre-load database in GSA system includes the GlobeLand30 dataset, geographic boundary data of 241 counties and their states or provincial boundaries. Thus, users can also define their ROIs through either the polygon tool of the selection in the pull-down menu of administrative divisions in the system interface. As discussed above, with the core data processing schemes build in GSA, for every user-requested ROI the system can provide the statistical analysis service that is calculated in a real-time manner and the same accuracy standard. Afterwards, the produced statistics results can also be exported as charts and tables saved in the local computer.

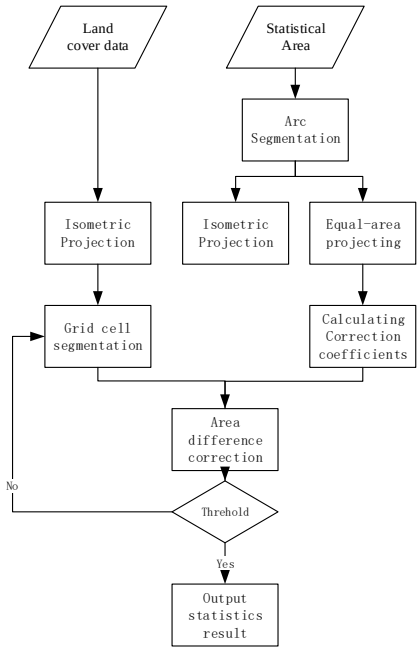


Figure 7. Flowchart of the developed strategy to correct the area distortion

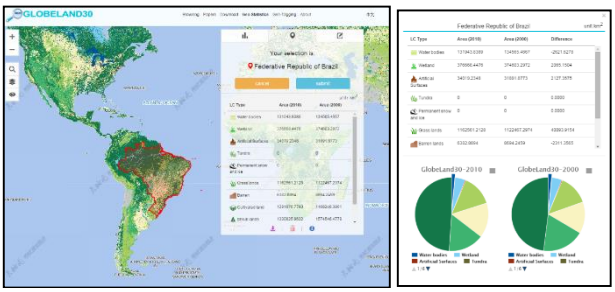


Figure 8 Demonstration of the user requests in GlobeLand30 GSA. (a) The user request and corresponding the land cover analysis result; (b) the analysis result of the land cover change between 2000 and 2010.

As shown in Figure 8, the web interface of GSA is demonstrated by an example request. In this example, Brazil is selected as the ROI, which is outlined by red polygon overlying on the GlobeLand30 map (Figure 8a). In the same figure, the statistics

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-4-2022-57-2022 | © Author(s) 2022. CC BY 4.0 License.

results in unit of square kilometres of ten land cover classes of Brazil in 2010 and 2000 are listed. These statistical results from two years can also be exported and save in a text file. Based on the land cover analysis information of these two different years, GSA can further provide detailed analysis of land cover change between 2000 - 2010 of every individual class (see Figure 8b). The land change analysis of this example request is shown in Figure 8b. In the real GSA interface, the request of conduct the change analysis can be done by clicking the information button below the statistics result in Figure 9a. Overall, the GSA has provide a clear, concise and friendly interface, for all user groups from different geo-informatics backgrounds. All output results can be saved in text files, thus, that can be used for any further analysis.

4. PERFORMANCE ASSESSMENT

In this section, we provide a study to evaluate the running performance of GSA online-statistic system, specifically the two aspect – the result accuracy and computation efficiency will be discussed here. The first test is focusing on evaluating the capability of our system to generate accurate land cover statistical analysis result; and the second experiment demonstrates the robustness of Globeland30 statistic system to provide quick response with large data set and multiple concurrent-requests condition. This performance assessment is conducted in order to further prove the benefit from the developed data processes algorithms discussed in Section 3. In the following discussion, together with the test result the details of test data as well as the implementation computation environment will be introduced, in order to fully understand the capability of the GSA online-statistic system.

4.1 Accuracy assessment

As shown in Figure 9, Sweden was chosen as the test site to evaluate the performance of the GlobeLand30 geo-statistics system in providing accurate result. Its latitude ranges from a wide value from 55 ℃N to 69 ℃N, whose statistics result is supposed to be greatly affected by different projection methods. Also, Sweden has a long coastline with the complicated shape that potentially introduces complexity in data processes.

4.1.1 Test data and implementation environment

The pre-loaded geographic boundary of Sweden was obtained from Digital chart of the World (DCW) dataset (Danko, et al., 1992) and its spatial reference is WGS 1984 geodetic system. As shown in Figure 10, seven maps sheets of GlobeLand30 data are needed to cover the whole Sweden test site. The corresponding index of map sheet is listed on the right side of the same figure.
The experiment was performed on a laptop with Inter Core I5-3230M CPU and 8GB Ram, the operation system is windows 7 professional. We also use the "Zonal Statistics as Table" in ArcGIS 10.1 desktop software to calculate the statistics of land use data for the comparison with the result from our algorithm. In this test, the vector data of Sweden's geographic boundary is projected into the cylindrical equal-area projection. It is also known as Behrmann projection. The area of the ROI calculated by vector data in equal-area projection is taken as the reference value of the ROI size.

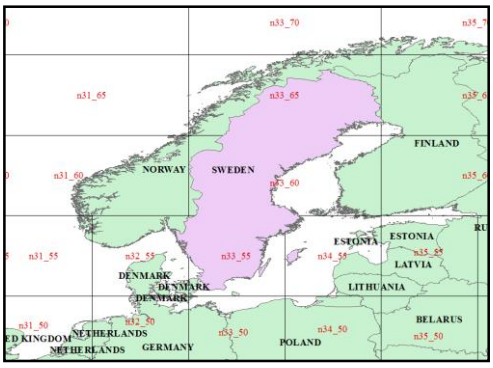


Figure 9. The geographic location of test site in the accuracy test. Light purple area outline the area of Sweden, which also shows that Sweden consists of seven map sheets.

4.1.2 Test result and discussion

ArcGIS has been used to generate statistic result for the same area for comparison. This is because ArcGIS is considered to be one of the most popular commercial geo-informatics analysis tools in both industry applications and scientific researches. Both the result of every land cover type from ArcGIS and our system are summarized and then valid with total area size of Sweden computed from DWG vector map after equal-area projection. We have used a relative error term (Eq.1) to quantify the errors from difference methods (ArcGIS, GlobeLand 30 statistic).

$$\xi_{relative} = \left| {A_{Ref} - A} \Big/ {A_{Ref}} \right| \qquad (1)$$

Where A_{Ref} is the reference area value of ROI, which is calculated by vector data in equal-area projection, and A is the area value calculated by land cover data from the two systems compared here. So that the relative errors from ArcGIS and the web-statistic system are considered as an accuracy indicator in this study. Additionally, we also record the elapsed time to process the same area in Globeland30 geo-statistics system and in ArcGIS, to evaluate the efficiency of developed algorithms in the web statistic system. In this study, the connection and transmitting time to the server through internet is considered, as it can vary at different internet condition (e.g., wireless, cable-line connection). Both the relative error and elapsed time are summarized in Table-1.
As shown in Table 1, it suggest that our algorithm implemented in Globeland30 online statistics system has lower relative error and consumed less compute time compared to the ArcGIS system, thus it has better accuracy and higher efficiency. Besides, comparing with the general GIS software, the GlobeLand30 geo-statistics system provides automated statistical analysis workflows, avoiding pre-processing work such as land cover data organizing, eliminating the overlap between adjacent map sheets of globeland30 data.
The statistic of every land cover class from both ArcGIS and the web-statistic system is shown in Table 1. Notice that given the size of Globeland30 grid cell is 30×30 m^2, the result value of each land cover type from ArcGIS tool is a multiple of 900 square meters. Whereas, the result values from GlobeLand30 are not, because our algorithms compensate the residual between isometric projection and equal-area projection. This also contributes the lower relative errors from our web-service discussed above.

	ARCGIS 10.1	OUR METHOD
STATISTICS RESULT	447,446,251,800.00	448,487,422,155.29
RELATIVE ERROR	0.36%	0.16%
TIME CONSUME	55.775	37.731
REFERENCE AREA	449,206,151,998.49	

Table 1. Result of the accuracy test

	ArcGIS	GlobeLand30 Statistic system
Cultivated land	38878416000.0	38914317928.10
Forests	250406163000.0	250479109067.93
Shrub Lands	909666000.0	909400512.28
Grass Lands	51044430600.0	51042957762.85
Wetland	29507071500.0	29508383859.71
Water bodies	37346339700.0	37363994993.81
Tundra	31710631500.0	31707607325.87
Barren Lands	5861936700.0	5866799649.30
Artificial Surfaces	1254043800.0	1256639947.44
Permanent snow and ice	527553000.0	527407766.56

Table 2. Result of the comparison between ArcGIS and Globeland30 statistics system for every class.

4.2 Computational efficiency and robustness assessment

As mentioned above in order to provide a user friendly web service, GSA has to be capable to provide quick response at various conditions, e.g., multiple concurrent user requests and/or large area request. Thus, we implemented a comprehensive analysis to evaluate the computational efficiency and robustness of the GSA to provide efficient service.

Two sub-tests have been designed to achieve this purpose. In the first one, we used the GSA web service interface to calculate the customized ROI request by users under the scenario of multi-users concurrent requests. The number of user request were submitted to the GlobeLand30 geo-statistics web system simultaneously. The number of request threads was increased from 10 to 200, thus the corresponding work load added to server's CPU and disk was also increased. The associated response time at different workloads has been recorded for the analysis. The relation between server response time and CPU utilization has been further analysed to evaluate the performance of the statistics system.

For the second sub-test, the purpose is to assess the robustness of the system. This is because as an online open service, the GSA has to continuously process massive requests and likely at the same time. This is to make sure the system would not crash if running in a long time. We used 50 threads request for geo-processing simultaneously, each threads kept 100 requests, resulting in 5000 requests in total. The CPU utilization and memory consumption were recorded for analysis indicators.

4.2.1 Test data and implementation environment

For the both efficiency and robustness test, the same ROI region is selected. It is a polygon region around the City of Guiyang in Guizhou province, China. This area is about 1650 square kilometres in size. The test was performed in a standard server within dual Xeon E5-2460 CPU at 2.6 GHz frequency and 16GB RAM, the operation system is Windows server 2008 64 bit. The Apache 2.4 was used as the HTTP server to provide the common gateway interface for executing the statistics program. Also, we prepared several python scripts to implement multi-thread requests, recording and monitoring the server's response and status.

4.2.2 Test result and discussion

The result of the concurrency request test can be visualized in Figure 11. It is used demonstrates the computational efficiency of the GSA system at the condition of multiple concurrency user requests. As it shows, when the concurrency number of user requests is more than 20, the server response time starts to increase. This is because the concurrency value is larger than the number of CPU logical cores, the request queue started to wait for executing. Increasing the number of concurrent requests, the difference between the average response time and maximum response time is getting more obvious. This is the result of the stacking order of processing instance from the request queue, which is hard to determine in the case of high concurrency. This experiment shows that although the response time of the GlobeLand30 geo-statistics system could be delayed in case of large amount of concurrent requests, while it is still capable of handling as many as 200 concurrent user requests and response in a bearable waiting time range.

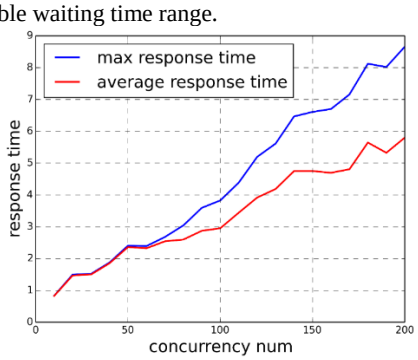


Figure 11. Result of concurrency test. The blue line denote the maximum response time of GSA and the red line denotes the average response time.

The result of robustness experiment is presented in Figure 12. By recoding the CPU utilization and memory consumption, we monitored the statistics system that ran 5000 requests. The request scheme is designed as to simulate the condition that 10 users sending out request at the same time and continues submit request one by one for total number of 500 requests. We record the usage of CPU and memory of the server in ever second, and this test took only 223 seconds in total. This result shows that the statistics system does not cost much server memory, since our algorithm only needs a small matrix block to mask land cover data. During this rest, the memory consumption had been kept at a level of 50%. And, in most times, the CPU utilization was between 60% and 70%, except for a few times when the geo-processing was initialized, the CPU utilization was getting 100%.

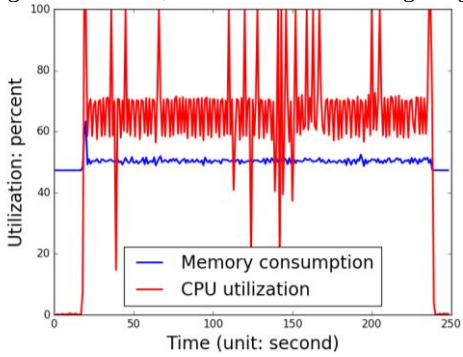


Figure 12 Result of robustness experiment. The x-axis is the elapsed time and y-axis is the percentage of usage of computer memory and CPU. The red line denotes the consumption of the

memory of the server system and blue line denotes the percent of CPU utilization.

4.3 Discussion

In the accuracy test, by comparing with the result from ArcGIS software, it suggest that the developed algorithms in GSA can provide statistical result with lower relative error and also consume less time compared to the Arcgis. Thus, the GSA has better accuracy and higher efficiency. Besides, the GlobeLand30 geo-statistics system provides an automated statistical analysis workflow. It also avoids pre-processing work from the user-end (such as land cover data organizing), eliminating the overlap between adjacent map sheets of globeland30 data.

In the efficiency experiment, GSA shows that although its response time could be delayed when there is a large amount of concurrent requests. However, it can still response in a bearable waiting time, when the number of concurrent user requests reaches 200. Thus, with the algorithms designed for GSA (see Section 3), the GSA remains sufficient system efficiency that is capable of handling as many as 200 concurrent user requests. In the robustness test, with 5000 requests and running for 250 seconds, the memory consumption has been kept at a level of 50%, and CPU utilization has been between 60% and 70% in most of the time. This is benefit from the dynamic matrix decomposition algorithms as we have discussed in Section 3.2.

Overall, through these two experiments on the result accuracy and computational efficiency and robustness, the GSA have shown a good performance. It can provide statistical results with high accuracy despite the geographic location, shape and complexity of requested, and can handle large number of concurrent user requests with sufficient efficiency and robustness. This performance is guaranteed by the data processing algorithms developed for GSA online-statistic system.

5. CONCLUSION AND FUTURE WORK

This paper introduces the online globeland30 GSA system and its key data processing strategies developed to generate statistics information with high accuracy from Globeland 30 data set. GSA is an on-line service provides analytic services with global land cover data. By optimally pre-organizing geography and auxiliary data base, GSA has multiple geospatial data processing modules and provides user an easy, and useful online service for global land cover data geo-statistics analysis. Running GSA does not need to install any special geospatial software in the local system nor require any geospatial information system skills and experience for users. Moreover, GSA is capable of generating accurate statistics analysis to user requests with various shape and size, and providing rapid response to multiple co-occurrence user-request.

In the future, more sophisticated geospatial analysis will be implemented to meet customers' requirements with growing amount and complexity, e.g., (1) providing portal for user uploading customized data layer(s), (2) providing more flexible and more diverse analysis capabilities by integrating multiple formats. Also, if the more detailed and up-to-date geographic boundaries data of individual administrative division can be available, the analysis result from GlobeLand30 geo-statistics system can be further improved. Additionally, this system's performance can be further improved by employing cluster servers or the cloud computing platform that can further enhance the processing efficiency that allows the system providing more complex analyses.

ACKNOWLEDGEMENTS

This work was supported by the Special Project of Science and Technology Basic Resources Survey, China Ministry of Science and Technology, under Grant 2019FY202502.

REFERENCES

Abbena, Elsa, Simon Salamon, and Alfred Gray. Modern differential geometry of curves and surfaces with Mathematica. CRC press, 2006.

Angel, Shlomo, Jason Parent, Daniel L. Civco, Alexander Blei, David Poter, The dimensions of global urban expansion: Estimates and projections for all countries, 2000–2050, Progress in Planning 75 (2011) 53–107 (IPF 1.75)

Bivand, Roger, and Albrecht Gebhardt. "Implementing functions for spatial statistical analysis using the language." Journal of Geographical Systems2.3 (2000): 307-317.

Brovelli, M. A., F. C. Fahl, M. Minghini, and M. E. Molinari. 2016. "Land Use and Land Cover Maps of Europe: A Webgis Platform." ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016: 913–917.

Cao X, Chen J, Chen L J, et al. 2014. Preliminary analysis of spatiotemporal pattern of global land surface water. Science China: Earth Sciences, 57: 2330–2339, doi: 10.1007/s11430-014-4929-x

Chen, J., Wang, D., Chen, Jun, et al. "Global land cover mapping at 30m resolution: A POK-based operational approach. 2015a ISPRS Journal of Photogrammetry and Remote Sensing 103 (2015): 7-27.

Chen Jun, Lijun CHEN, Ran LI, etal . 2015b. Spatial Distribution and Ten Years Change of Global Built-up Areas Derived from GlobeLand30[J] . Acta Geodaetica et Cartographica Sinica, 2015, 44(11):1181-1188

Chen, Jun, et al. "Towards a collaborative global land cover information service." International journal of digital earth 10.4 (2017): 356-370.

Han, G., J. Chen, J., C. He, S. Li, H. Wu, A. Liao, and S. Peng. 2015. "A web-based system for supporting global land cover data production." ISPRS Journal of Photogrammetry and Remote Sensing 103: 66-80.

Haining R. Spatial Data and Statistical Methods: A Chronological Overview[M]//Handbook of Regional Science. Springer Berlin Heidelberg, 2014: 1277-1294

De Jesus, Jorge, P. Hiemstra, and G. Dubois. "Web-based geostatistics using WPS." Proceedings of the 6th Geographic Information Days 32 (2008): 199-218.

Han, Weiguo, et al. "CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support." Computers and Electronics in Agriculture 84 (2012): 111-123.

Pettit, C. J., Tanton, R., & Hunter, J. (2017). An online platform for conducting spatial-statistical analyses of national census data

across Australia. Computers, Environment and Urban Systems, 63, 68-79.

Chen, J., Ban, Y., Li, S., 2014. China: open access to Earth land-cover map. Nature, 514 (434), 23.

Danko, David M. "The digital chart of the world project." Photogrammetric engineering and remote sensing 58.8 (1992): 1125-1128.

Fernandez, Carlos J., and T. Neal Trolinger. "Development of a Web-Based Decision Support System For Crop Managers." Agronomy journal 99.3 (2007): 730-737.

Mora B, Tsendbazar N E, Herold M, et al. Global land cover mapping: Current status and future trends [M] //Land Use and Land Cover Mapping in Europe. Springer Netherlands, 2014: 11-30.

Ramankutty, Navin, and Jonathan A. Foley. "Characterizing patterns of global land use: An analysis of global croplands data." Global Biogeochemical Cycles 12.4 (1998): 667-686.

Scott, G., and A. Rajabifard. 2015. “Integrating Geospatial Information into the 2030 Agenda for Sustainable Development.” Paper presented at the 20th United Nations Regional Cartographic Conference for Asia-Pacific Jeju Island, Republic of Korea, October 6-9.

Tateishi, Ryutaro, et al. "Production of global land cover data–GLCNMO." International Journal of Digital Earth 4.1 (2011): 22-49.

Townshend J R, Masek J G, Huang C, et al. Global characterization and monitoring of forest cover using Landsat data: opportunities and challenges[J]. International Journal of Digital Earth, 2012, 5(5): 373-397.

Whigham, Dennis F. 2009, Global Distribution, Diversity and Human Alterations of Wetland Resources, The Wetlands Handbook Edited by Edward Maltby and Tom Barker, © 2009 Blackwell Publishing Ltd. ISBN: 978-0-632-05255-4

Yang, Chaowei, et al. "Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing?." International Journal of Digital Earth 4.4 (2011): 305-329.

Zell, Erica, et al. "A user-driven approach to determining critical earth observation priorities for societal benefit." Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of 5.6 (2012): 1594-1602.