ACCURACY ASSESSMENT OF GLOBELAND30: A CASE STUDY OF ONTARIO, CANADA

by

Rishita Rangarh

B.Tech., University of Petroleum and Energy Studies, Dehradun, India, 2014

A MRP

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Engineering

in the Program of

Civil Engineering

Toronto, Ontario, Canada, 2016

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MRP

I hereby declare that I am the sole author of this MRP. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

ACCURACY ASSESSMENT OF GLOBELAND30: A CASE STUDY OF ONATRIO, CANADA

Master of Engineering 2016, Rishita Rangarh, Civil Engineering 2014-2016

Abstract

GlobeLand30 is the world's first 30m high resolution land cover data set (Chen et al. 2014) and has been a successful model of Big-Data mining from a host of Landsat imagery, thereby contributing to and enhancing the existing global geospatial knowledge base (GlobeLand30 2014). As there is a lot of uncertainty and errors in the global land cover data, therefore it becomes very difficult to validate land cover on a global scale. Efforts on validating Globeland30 data have been made in various parts of the world in the past and will continue to be done. The objective of this project is to validate GlobeLand30 data set by carrying out a case study in Ontario, Canada. The adopted methodology for doing validation is by using cell-to-cell benchmarking (Maria et al. 2015), thereby deriving Error Matrix, and its derivatives, which includes overall accuracy, user accuracy, producer accuracy and kappa coefficient. The results show that an overall accuracy of 84.14% is obtained for GlobeLand30 data with consideration of shadows, which is relatively a high percentage number indicating that the GlobeLand30 data classification is highly accurate for Ontario, Canada.

Keywords: land cover; GlobeLand30; accuracy assessment; Ontario

Acknowledgements

First, I would like to thank my supervisor Dr. Songnian Li for his able guidance and incessant support throughout my degree. I am deeply grateful to him for supporting, inspiring and always being there when I needed him. He provided me with all the necessary resources that enhanced my knowledge and a very comfortable research environment. I learned a lot under his supervision and I consider myself lucky to have such a great opportunity to work on this project as it is conducted for the first time on a Canadian landscape. While working on this project I learned a lot about GIS software handling and some life changing lessons for which I solely thank Dr. Li and this university for giving me a chance to hone my skills. I would also like to extend my appreciation to our Graduate Administrator Rachel Peluso as she has been very patient and so approachable throughout my tenure at Ryerson. She is a supportive and understanding person and we as International Students have so many apprehensions and queries, which she used to answer always without fail. I owe my sincere gratitude to Dr. Ahmed Shaker who taught me Remote Sensing and made it so interesting, I am sure it will definitely help me in my career.

I am highly indebted to have a number of professors in the department who gave me guidance and taught me skills that would not just help me in my professional career but also in my personal life. I would like to express my deepest gratitude to Dr. Ahmed Al Rabbany who was so kind and helpful even when I did not take any courses under him. Thank you professor for all your support, and best wishes.

Last but definitely not the least I would like to thank my parents because of whom I made it to this University. Thank you for always believing in me and providing me such a loving and open environment where I could be myself and grow. My deepest gratitude to my father who has been my pillar of strength and support system in all walks of life. I love you! Big thank you to Srishtie Srivastava because of whom there was never a dull moment in my life here.

Rishita Rangarh

Table Of Contents

List Of Tables

List Of Figures

1. Introduction

Land cover information has become an integral part for a myriad of societal needs ranging from environmental studies, natural resources management, urban planning to sustainable development (Verburg et al. 2010). Land cover refers to a combination of various physical material types occurring on the surface of the earth along with their attributes and features (Feddema et al. 2005). The Ministry of Science and Technology of China sponsored a National High Technology Research and Development program under its 863 program entitled "remote sensing mapping and research on key technologies of global land cover" in 2010 and one of the outcomes of that project was GlobeLand30- 2000 and 2010 data sets (NGCC 2014). It is a global land cover-mapping product derived from remote sensing images mainly 30m multispectral images, including Landsat TM and ETM+ multispectral images, Chinese Environmental Disaster Alleviation Satellite (HJ-1) multispectral images (GlobeLand30 2014). Further, ancillary data such as regional and national land cover maps, MODIS NDVI data and global elevation data for the year 2010 at a 30-meter spatial resolution (Duhok-Globeland30). These detailed data sets will assist researchers and scientists from across the globe to better understand, monitor and detect changes in land cover and land use all over the planet (Foody et al. 2011; Olofsson et al. 2014; Stehman et al. 1998; Strahler et al. 2006). The land area covered by these data sets extends from 80 degree North to 80 degree south and is classified into ten land cover types, namely, cultivated land, forest, grassland, shrub land, wetland, water bodies, tundra, artificial surfaces, bare land, permanent snow and ice (NGCC 2014). The positional accuracy of GlobeLand30 data sets is +/- 75 meter. The data sets adopt the WGS84 coordinate system with UTM projection. Minimum Mapping Unit (MMU) has been adopted to assess the quality control of these data sets by taking into consideration the smallest patch and if the area patch is larger than the MMU size, it has to undergo the quality control process. The extraction is done using a hierarchal extraction method where each land cover type is one by one classified along with the constraints of the mask of other land cover types. Figure 1 illustrates the POK based approach that has been used for operational mapping.

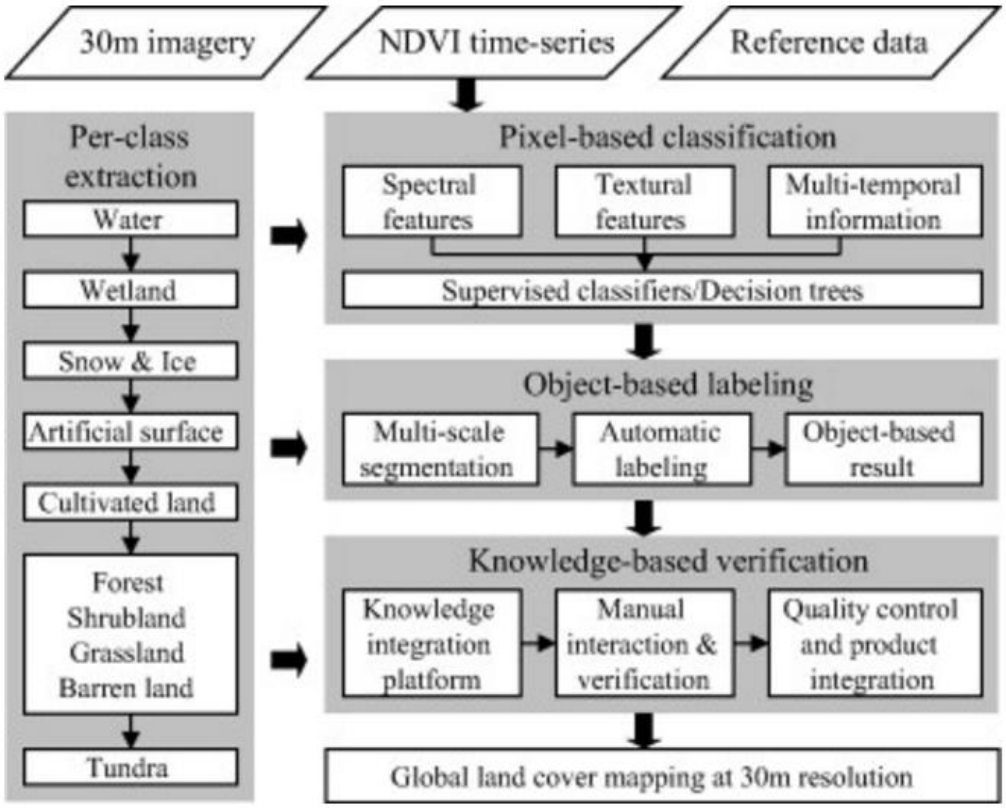


Figure 1 The developed POK classification approach (source: NGCC 2014)

GlobeLand30 product is available in both vector (.shp) as well as raster (geoTIFF) formats, however, for this project only raster data is considered for performing accuracy assessment. Figure 2 below illustrates the accuracy of each type of land cover in GlobeLand30-2010.

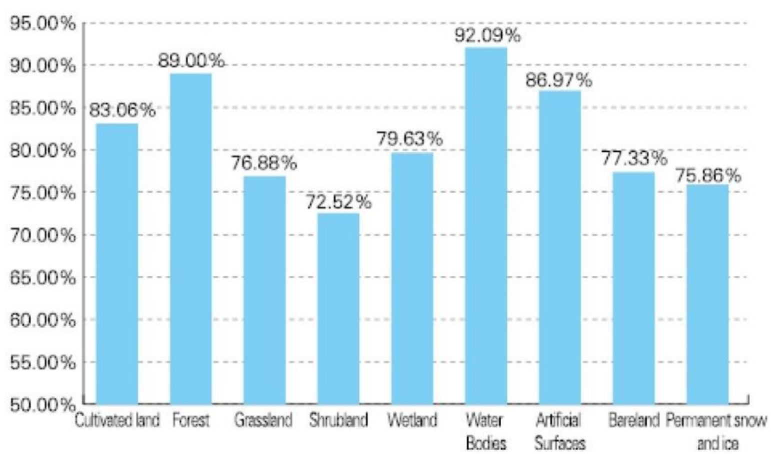


Figure 2 Accuracy of each Type in GlobeLand30-2010 (source: NGCC 2014)

Validating data is one of the key issues and tasks in global land cover mapping, as it is very important to know the level of quality of the data we use for data analysis and decision-making. Validation plays a key role in the land cover mapping because without validating the classification against higher-quality reference data, any land cover map would just be a mere untested hypothesis and is of no use to any research organization or academic sector (Maria et al. 2015). If the data we are using is of acceptable range, it builds a sense of confidence to use the desired data for carrying out projects. Validation helps in detection of anomalies with features, attributes, and relationships in the database (ESRI 2016). Due to the fact that validation is not graded, we have to be very cautious during validation. Even though the process of validation is automatic but it consists of a lot of labour work. Therefore, most global land cover maps are either cross validated from training samples or estimated with a limited number of samples mainly hundred or thousand (ground control points) (Figure 3)

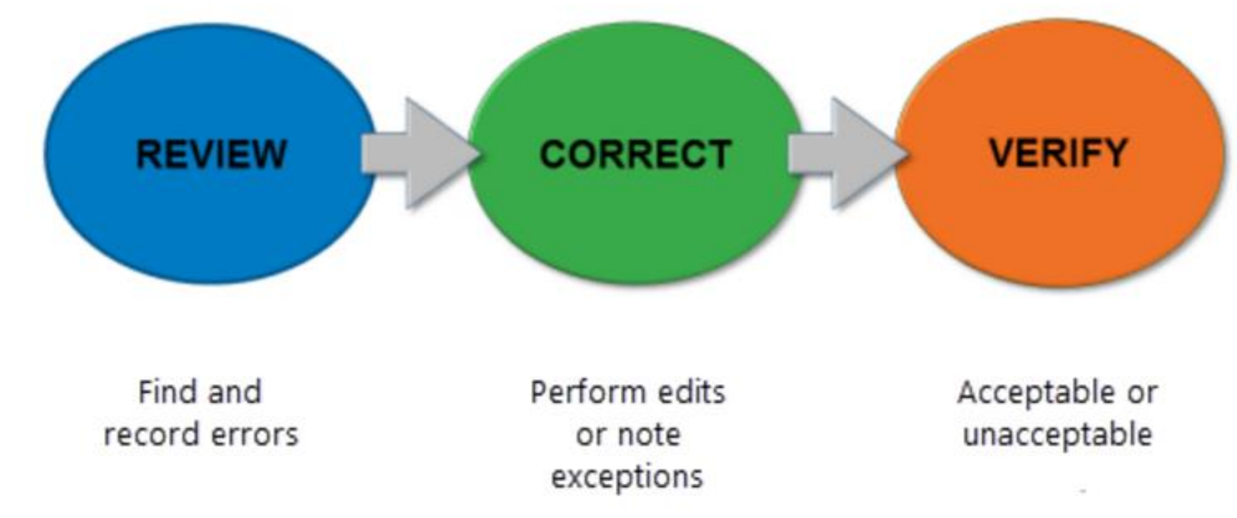


Figure 3 Validation Flowchart (source: ESRI 2016)

In order to validate GlobeLand30 data many studies in the past have been carried out. However, at a National level apart from Ontario, Canada - which is the area of study for this project, similar assessment was been carried out for some regions in Italy, Germany and some parts of Central Asia. In Germany, a study was conducted to compare GlobeLand30 with existing data set (i.e., CORINE, Urban Atlas, OpenStreetMap, and ATKIS) Jamal et al. (2016). The purpose of this study was to compare GlobeLand30 with existing authoritative and crowdsourced Land cover products for Italy and Germany in order to evaluate the agreement. In Italy, the study performed was the first accuracy assessment at a National level and the methodology adopted was benchmarking with the more detailed Italian land cover data sets (Maria et al. 2015). The overall accuracy value for Italian data sets turned out to be higher than 80%. Another study to validate GlobeLand30 data set was carried out in Central Asia, located in the hinterland of the Eurasian Continent (Sun et al. 2016). Investigations conducted by KTH Royal Institute of Technology Sweden and the Institute for Applied Systems Analysis (IIASA) in Northern Europe shows that the accuracy of GlobeLand30 (total disagreement less than 5%) was much higher than the other existing GLC products (GlobeLand30 2014).

This study mainly focuses on assessing the accuracy of GlobeLand30 classification for Ontario cell-by-cell comparison is carried out for benchmarking and further, re-projection, tile merging and reclassification is carried out to derive the accuracy measurement indicators that includes Error Matrix and its derivatives namely Overall accuracy, User accuracy, Producer accuracy and Kappa Coefficient (Maria et al. 2015).

The report is organized as follows. Section 2 illustrates the adopted methodology that includes benchmarking between GlobeLand30 data set and Ontario land cover data set. It further includes description of data set for both GlobeLand30 as well as Ontario land cover. Section 3 reports the results and discussions supported by statistical parameters. Finally, the conclusion in Section 4 illustrates the main objective around which the whole project is revolving.

2. Methods and Data

2.1 Confusion Matrix and Derived Indexes

In order to validate the classification quality of thematic maps, methods such as creating our own random sample points, ground control points that is mainly our reference of doing the validation of a classified image. As, this project is a case study for Ontario region of Canada, the methodology followed is similar to the accuracy assessment done for Italian territory (Maria et al. 2015). In this study, accuracy assessment of GlobeLand30 was carried out through comparisons (Maria et al. 2015) with an already existing data sets of Ontario, Canada. However, there is a set of practices that is described in the following paragraph that one needs to keep in mind before going ahead with the validation and those practices are called "good practices" suggested by Foody et al. (2014) and Olofsson et al. (2014)

Planning accuracy assessment of reference data that is land cover data set for the province of Ontario; was done using a set of criteria that involves constructing an efficient and practical statistically rigorous model based on probability-sampling and consistent estimation. The three widely accepted basic components for performing accuracy assessment are: 1) the sampling design used to "select the reference sample"; 2) the response design used to obtain the reference land-cover classification for each sampling unit; and 3) "the estimation and analysis procedures" (Stehman et al. 1998).

Response Design: It refers to the schema/workflow followed for the acquisition of the reference data set (Foody et al. 2011) and the key factors in response design includes spatial assessment unit (Stehman et al. 2011), source of reference data and the reference labels. The classification quality of the reference data set should be much higher than the data set that is chosen for validation as only then a better understanding of the results could be predicted and analyzed. It involves procedure for gathering information pertaining to the referenced land cover determination and methods for assigning more than one or sometimes just a single classification to each sampling unit. Accuracy is determined by comparing land cover map classification to the referenced classification and determining the extent of agreement between them.

Sampling Design: The basis of accuracy in this system is established by defining a sampling frame along with the sampling unit. This step comes into play if the collection of reference data set is not possible then the probability sampling design is created to collect the sample data on which the accuracy validation can be performed. Hence, choice of both sample size and sampling methods should be very accurate or else it will not be of much use and will not be able to test the accuracy precisely (Stehman et al. 2009).

Accuracy analysis: This step is an extension to Response Design and Sampling Design because if the above-mentioned two steps are coordinated, then comprehending accuracy becomes a lot easier.

For the present study, reference data sets were selected among the already existing data set for the province of Ontario, Canada. The data set taken into consideration is the Ontario provincial land cover map that provides a classification of 27 broad land cover types across province of Ontario generated for the year 2000.

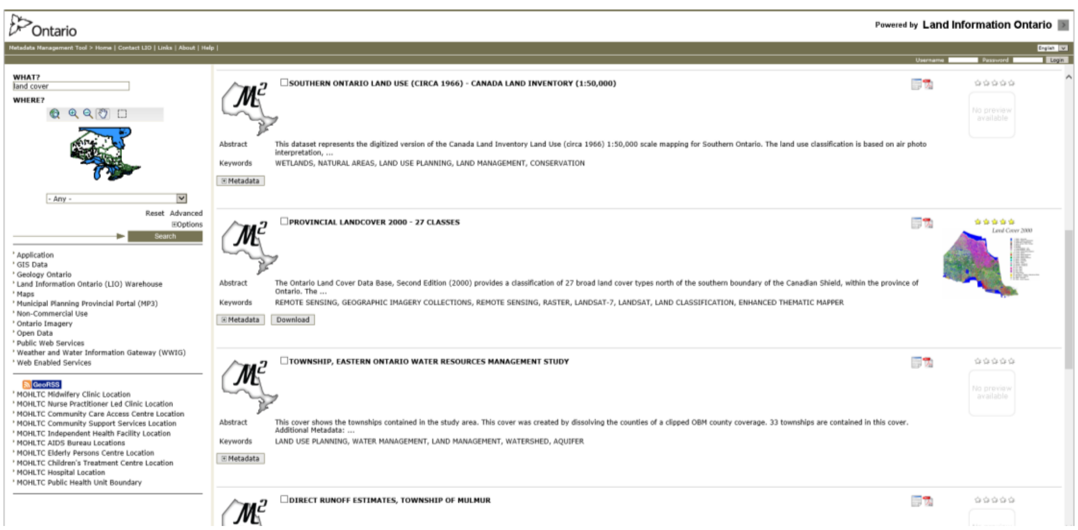


Figure 4 Data source for Ontario provincial land cover class map- year 2000

Coming on to the accuracy assessment, cell-by-cell comparison is done between the classified and the reference data set by selecting the pixel as the unit for doing benchmarking and this gives us an error or confusion matrix (Congalton et al. 1999) and its derivatives. The error matrix is a square matrix consisting of number of rows and columns depending on the total number of land cover classes been considered for the assessment. Error matrix and kappa coefficient helps in quantifying the data quality by depicting the classification accuracy at pixel level. Quality indicators include overall accuracy, producer's accuracy and user's accuracies (Liu et al. 2007). Equation 1 was used to calculate Kappa Coefficient (Pontius 2011):

$$Kappa = \frac{N\ \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} (x_{i+} . x_{+i})}{N^2 - \sum_{i=1}^{r} (x_{i+} . x_{+i})} \tag{1}$$

Where,

r = number of rows in error matrix

x_{ii} = number of observations in row I and column i

x_{i+} = total of observations in row i

x_{+i} = total of observations in column i

N = total number of observations included in matrix

Overall Accuracy (OA) is one of the most used agreement measures and it calculates the percentage of correctly classified pixels. In theory, it is next to impossible to find an ideal threshold value for overall accuracy. However, Anderson et al. (1976) came up with a value of at least 85%, Pringle et al. (2009) with a value over 70%, whereas Thomlinson et al. (1999) said that the classification is accurate if the OA is equal to 85% with no less than 70% accuracy of each class. Equation (2) stated below was used to calculate OA (Olofsson et al. 2014)

$$\frac{Number\ of\ test\ samples\ correctly\ classified\ (Sum\ of\ diagonal\ elemnts)}{Total\ number\ of\ test\ samples\ (Sum\ of\ rows\ in\ a\ confusion\ matrix)}$$

(2)

Producer accuracy and user accuracy deals with the accuracies of individual classes unlike overall accuracy. Producer Accuracy gives the accuracy measurement from the producer perspective and is defined as the percentage of the pixels correctly detected in a classified map (Olofsson et al. 2014).

Producers Accuracy (Omission errors):

$$\frac{Number\ of\ samples\ of\ a\ class\ correctly\ classified}{Number\ of\ samples\ of\ the\ class} \tag{3}$$

Whereas, user accuracy gives the accuracy measurement from the user's perspective and is statistically defined as the percentage of pixels that actually belongs to a class (Olofsson et al. 2014).

User's Accuracy (Commission errors):

$$\frac{Number\ of\ samples\ that\ actually\ belong\ to\ a\ class}{Number\ of\ samples\ assigned\ to\ that\ class} \tag{4}$$

If we need to derive practical decisions regarding image classification then knowing the value of kappa is not sufficient and thus, we need to know the value of proportion correct. Therefore, Allocation Disagreement (AD) and Quantity Disagreement (QD) which are relatively new parameters proposed by Pontius et al. (2011) are calculated for estimating the disagreement component between classified and reference data set. When the disagreement is, "due to the less than optimal match in the spatial allocation of the categories" it is termed as allocation disagreement and when it is "due to the less than perfect match in the proportion of the categories" it is termed as quantity disagreement (Pontius et al. 2011). Equation (5) and (6) defines the formula to calculate AD and QD:

$$AD = \frac{\sum(2 * \min(\frac{n_{+i}}{n} - \frac{n_{ii}}{n}, \ \frac{n_{i+}}{n} - \frac{n_{ii}}{n}))}{2} \times 100$$

(5)

$$QD = \frac{\sum(\frac{n_{+i}}{n} - \frac{n_{i+}}{n})}{2} \times 100$$

(6)

Where,

$n =$ total number of considered pixels

$n_{ii} =$ diagonal elements

$n_{i+} =$ marginal sum of the rows

$n_{+i} =$ marginal sum of the columns

2.2 Available Data

2.2.1 Globeland30

GlobeLand30 is by far one of the first global land cover data set at 30meter resolution. It is the result of the project entitled “Global land Cover Mapping at Finer Resolution”, led by the National Geomatics Center of China (NGCC). The GlobeLand30 data sets are available for two years: 2000 and 2010 and produced over a period of four years (GlobeLand30 2014). In line with NGCC's aim to help the global community (NGCC 2014) in assessing the accuracy of this product in different parts of the world, this project attempts to study and compare the GlobeLand30 data set (Chen et al. 2014) with the current available sets that exist for Canadian LC mapping. The land area covered by the GlobeLand30 data set extends from 80 degree North to 80 degree South and is classified into ten land cover types, namely, “ cultivated land, forest, grassland, shrub land, wetland, water bodies, tundra, artificial surfaces, bare land, permanent snow and ice (NGCC 2014)”. The positional accuracy of GlobeLand30 data set is +/- 75 meter (Chen et al. 2014; Jokar et al. 2015, 2015a, 2015b, 2013, 2014; Jokar and Vaz 2015; Vaz et al. 2015). This data set adopts the WGS84 coordinate system with UTM projection. Minimum Mapping Unit (MMU) is adopted to assess the quality control of these data set by taking into consideration the smallest patch and if the area patch is larger than the MMU size then it has to undergo the quality control process (NGCC 2014). The extraction is done using a hierarchal extraction method where each land cover type is one by one classified along with the constraints of the mask of other land cover types (NGCC 2014). The POK approach for operational mapping is used. The GlobeLand30 project was initiated by the National Geomatics Center of China (NGCC) (GlobeLand30 2014). Figure 5 illustrates the GlobeLand30 data in raster format for southeast and southwest region in the World Map and similarly we have for northeast and northwest. The raster data is in the form of tiles and for collection and classification, more than 10,000 scenes were obtained and is considered to derive Globeland30 data. The images considered for data generation of GlobeLand30 were mainly multispectral images including Landsat TM and ETM+ multispectral images and multispectral images of Chinese Environmental Disaster Alleviation Satellite (HJ-1) (NGCC 2014).

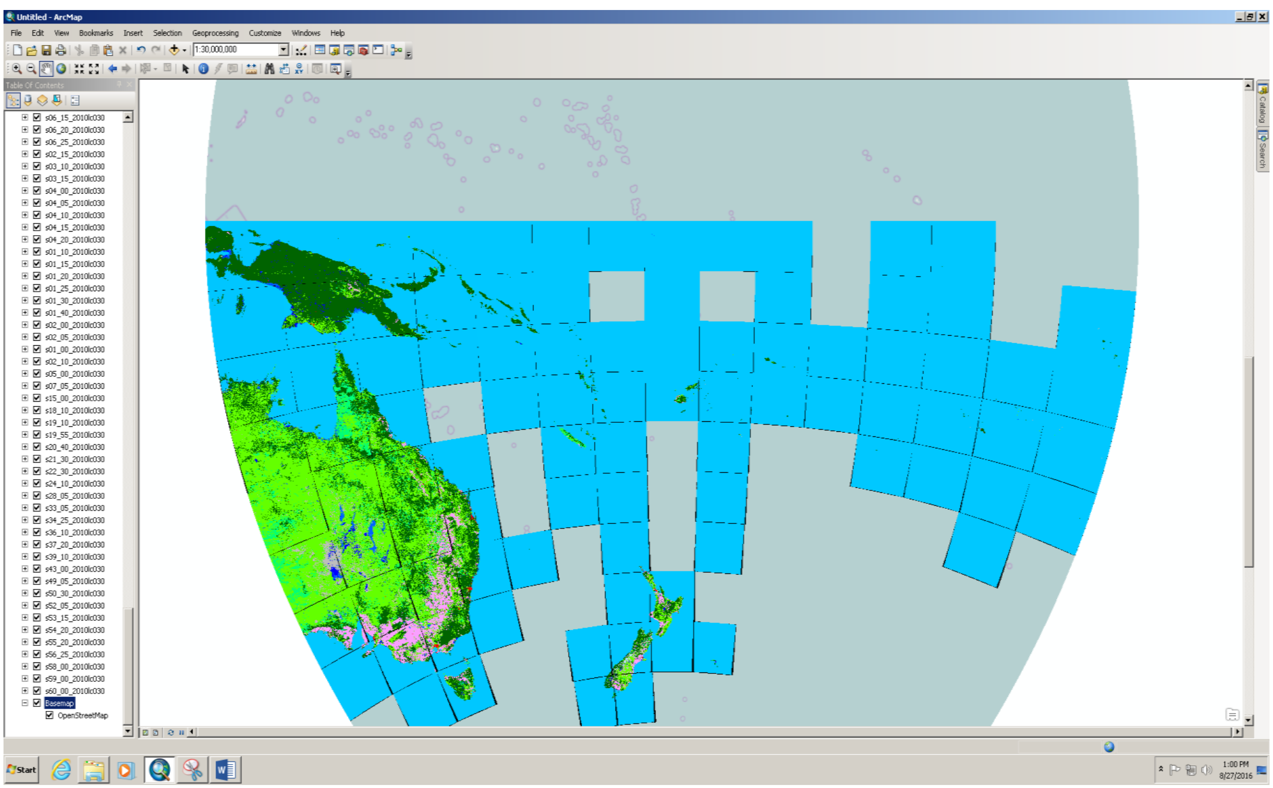


Figure 5 Overlay of tiles for GlobeLand30 for the southern part of the globe, year-2010

For vegetation growing season in particular cloudless images were selected with a +/- of one year from 2010 and if there were missing images of any area, the time frame was extended (NGCC 2014).

Figure 6 below illustrates the GlobeLand30 map for the baseline year 2010 (NGCC 2014)

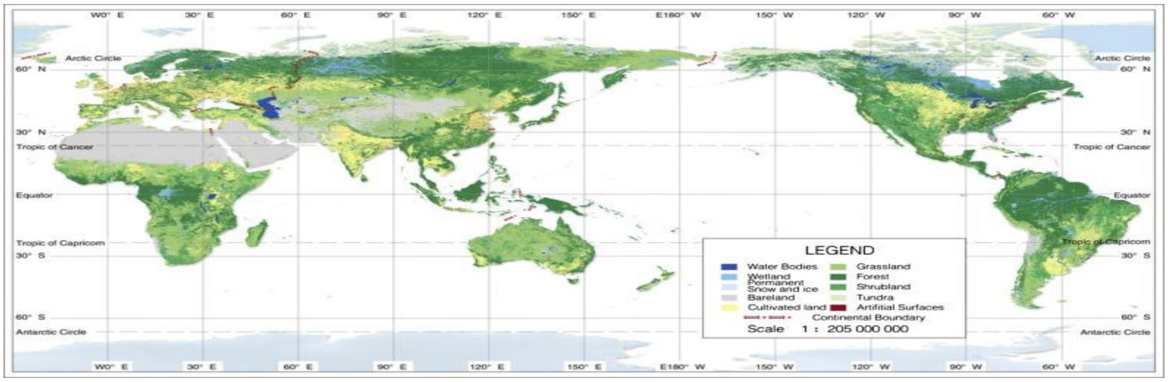


Figure 6 GlobeLand30 map for the year 2010

The open access of this important and detailed scientific data set will significantly promote scientific data sharing and assisting in deducing conclusions in the field of earth observation and geospatial information sciences within the international community (NGCC 2014). GlobeLand30 data for the present study (Ontario, Canada) is available in raster format with 12 different tiles covering the Ontario region of Canada. The data is in WGS84 (World Geodetic System 1984) reference system and UTM (Universal Transverse Mercator) projection (NGCC 2014). The map below illustrates GlobeLand30 for the Ontario territory and the legend is based on ten land cover classes (NGCC 2014).

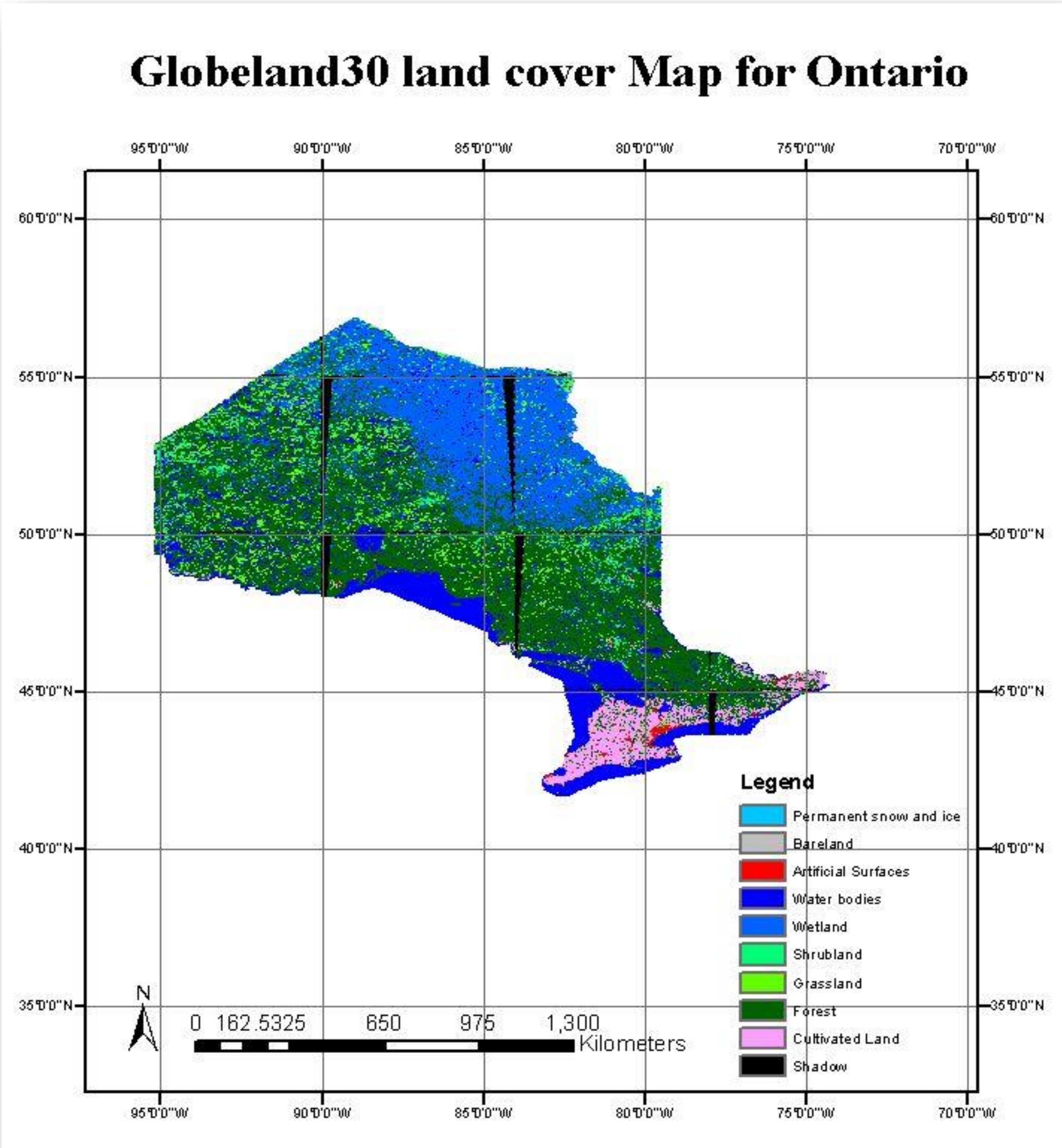
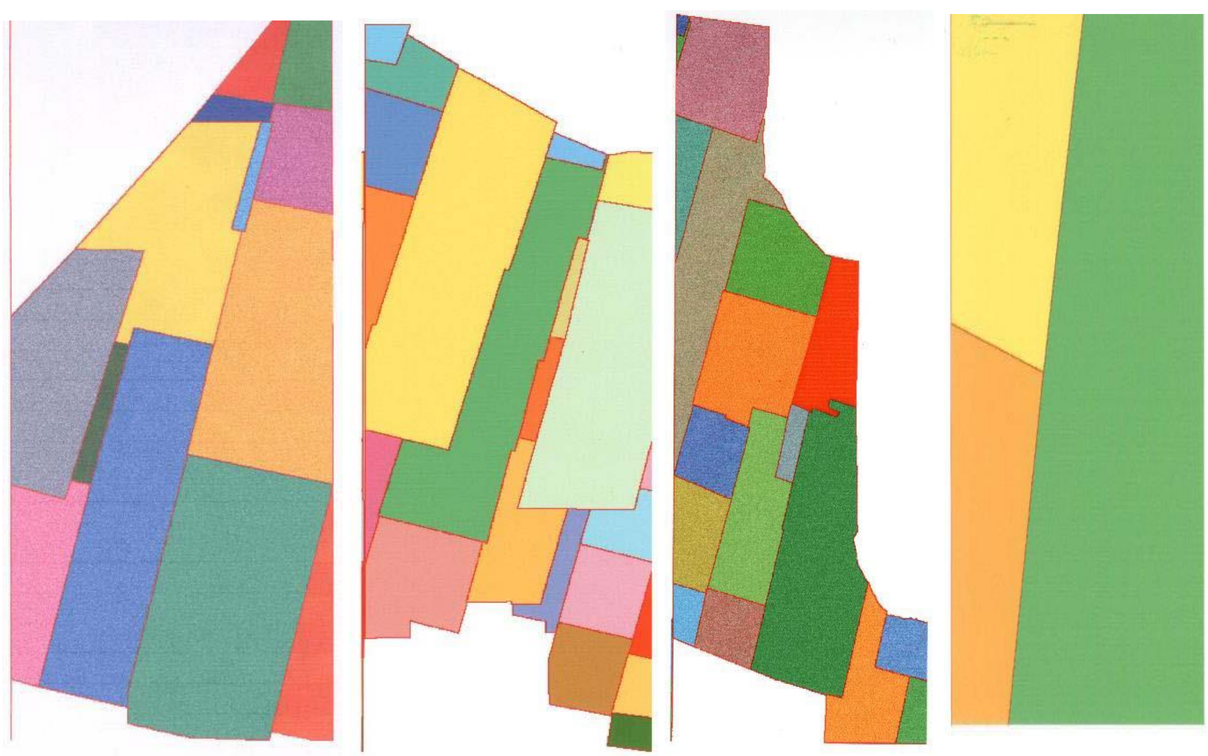


Figure 7 GL30 Land Cover map

2.2.2 Ontario Land Coverage data set

The reference data acquired for the current study to carry out accuracy assessment of globeland30 data set is the Ontario Provincial Land Cover Data classified into 27 broad land cover types across the province of Ontario for the year 2000 obtained from the Ontario portal for open data set (Spectranalysis Inc. 2004).

The data is generated from Landsat-7 Thematic Mapper (TM) satellite data frames taken in between the year 1999 and 2002, mostly from 2000 onward. For the land cover classification source images considered were ortho-rectified frames provided by Natural Resources Canada. The reference data set for the province of Ontario is also available in raster format with four different tiles covering the Ontario region of Canada. The land cover database extends over UTM Zones 15, 16, 17 and 18. Further, the data set scale range is 1:50,000 and the data available is in North American Datum (NAD1983) reference system (Spectranalysis Inc. 2004).



(a) (b) (c) (d)

Figure 8 Ontario Land Cover database UTM Zone 15 (a), UTM Zone 16 (b), UTM Zone 17 (c), UTM Zone18 (d)

All the data set are available with accuracy equal to 25m and the minimum mapping unit is eight-pixel. Finally, the best part about the data set is that it agrees with the CLC nomenclature, which is a hierarchical classification system, based on three levels (Spectranalysis Inc. 2004) and in this study, we will be taking into account whose first level comprises of five classes namely: artificial surfaces, agricultural areas, forests and semi natural areas, wetlands and water bodies.

Figure 9 illustrates the reference data set: Province of Ontario land cover Map based on 27 land cover classes.

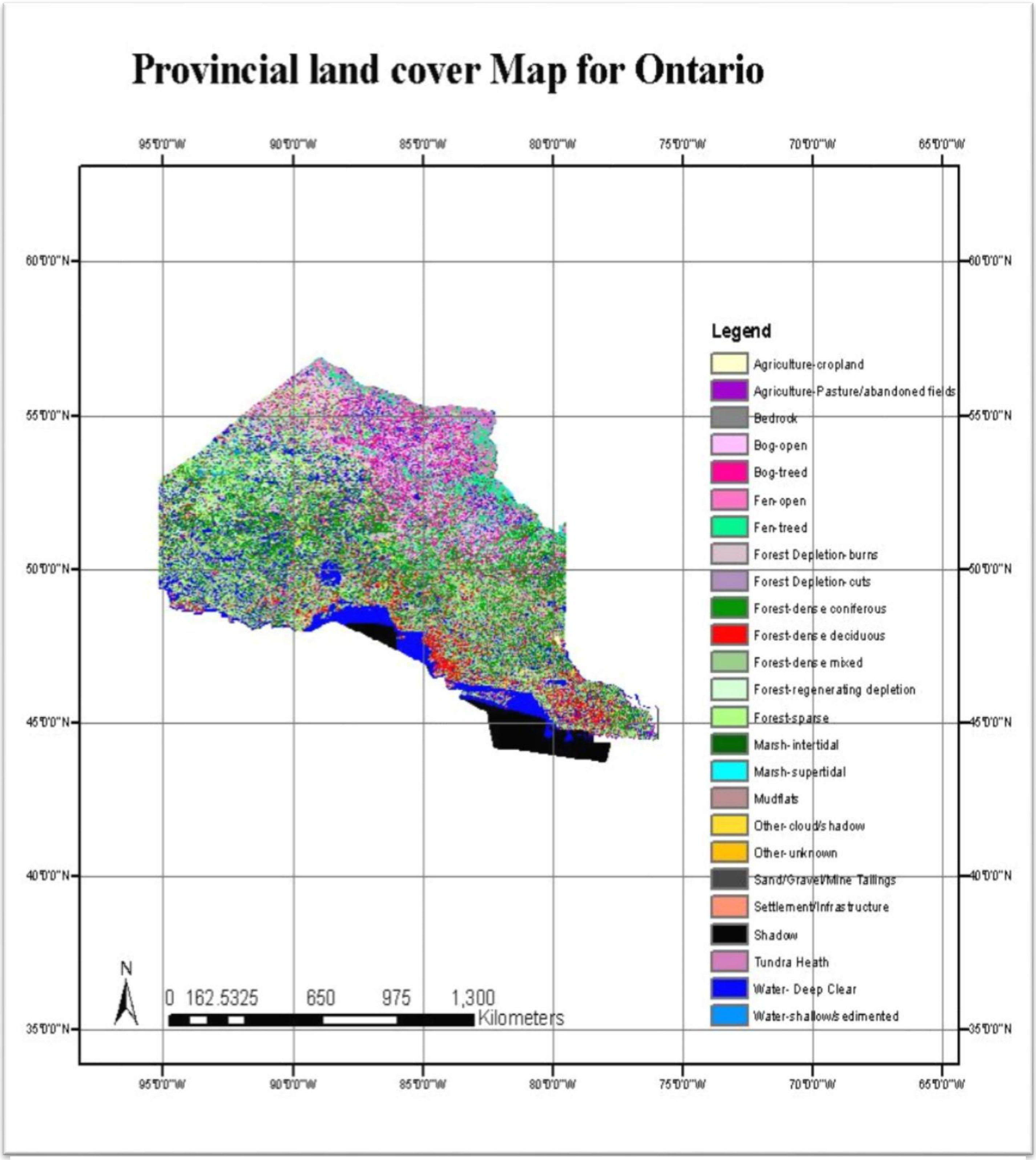


Figure 9 Ontario Land Cover Map (source: ArcGIS)

2.3 Data Processing

As GlobeLand30 and the Ontario regional maps have different legend, scale and reference system thus, a data pre-processing (Maria et al. 2015) phase is very essential before going ahead with calculating confusion matrix and its derived statistics. The schema followed for data processing is shown in Figure 11. However, for the current study we skip the step Rasterization (converting from vector to raster) (Biagi et al. 2015) in the Region section of the flow chart as the data for the province of Ontario is already in the raster format (Spectranalysis Inc. 2004) and thus rasterization is not needed. Re-projection is another very important step because if both the data sets are not in the same projection system then it becomes difficult to draw a comparison and also overlaying the two maps is not possible (Maria et al. 2015). Hence, both the data set are re-projected to WGS84 (World Geodetic System 1984) reference system. The number of tiles available for each data set namely for GlobeLand30 we have 12 tiles covering the Ontario region and for the reference data set we have four tiles that gives us the entire provincial land cover. Thus, mosaicking (Xu et al. 2010) of these tiles is performed and after applying mosaic, a single merged tile for each data set is obtained. Next step is to apply raster editing on these Mosaic tile obtained for each data set and by raster editing, it means to clip the Mosaic with respect to the polygon boundary of Ontario (Scholars geoportal 2015)

Figure 10 Ontario boundary selected and data sets extracted from Scholars Geoportal (source: Scholars geoportal 2015)

This has been done so that the extent of both the maps are the same so that accuracy analysis can be done precisely and the comparison becomes easy and attainable. Further, if the extent of the maps taken for study is not the same or the difference is huge then it might result in altering the value of the classified classes taken for comparison and thus resulting in errors largely. Once the Mosaic for both GlobeLand30 and Ontario region is clipped with respect to the Ontario provincial boundary, the data sets are ready for the final and the most important step Reclassification. See Figure 11 for the overall workflow.

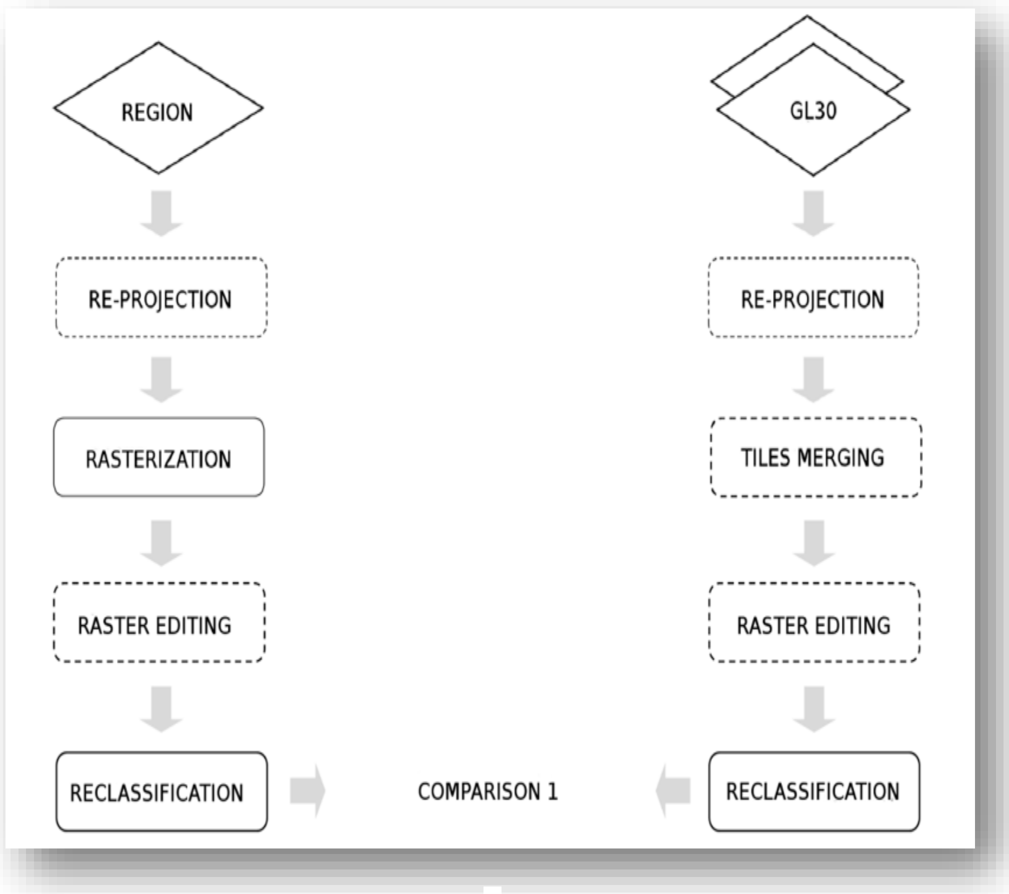


Figure 11 Data Processing workflow performed on Ontario data set and GlobeLand30 (GL30) (source: Maria et al. 2015).

Reclassification is another very important step as no comparison is possible between GlobeLand30 and Ontario data sets because it is characterized by different thematic classification. Thus, the reclassification method selected for reclassifying both the maps is reclassification based on first five level of CLC (Corine Land Cover) nomenclature that means that it consists of five land cover classes namely: artificial surfaces, agricultural areas, forests and semi natural areas, wetlands and water bodies. The reason behind selecting the method with first five level of CLC is that it allows the overall accuracy to be between 81% and 92% while the second method is based on the wide availability of GlobeLand30 for classes for forests and semi-natural areas. Thus, replacing these classes with four sub-classes namely forest, grass and/or herbaceous vegetation associations, open spaces with little or no vegetation, glaciers and perpetual snow; allows an overall accuracy to result somewhere in between 62% and 81% Maria et al. (2015). Therefore, first reclassification method has been selected for this study and the reclassification is carried out manually because the same methodology for reclassification (Maria et al. 2015) has been followed. In ArcGIS, the reclassify tool in the spatial analyst toolbox is used for performing reclassification in which both the images are given new legends based on five land cover classes of corine legend. The translation is done by changing the value of the "value" attribute for every class in the non-reclassified image (Refer Section 3.3.1 for more details). It is important to note here that the probability of error is high because of the ambiguous interpretation of classes done manually by me. Table 1 depicts the correspondence between GlobeLand30 and Corine classes (Maria et al. 2015).

Corine Legend	GlobeLand30 Legend
1 Artificial surfaces	1 Artificial cover
2 Agricultural areas	2 Croplands
3 Forests and semi natural areas OR	3 Mixed forest, Broadleaf forest, Coniferous forest, Grass, Shrub, Bare land, Permanent ice or snow OR
3.1 Forests 3.2 Scrub and/or herbaceous vegetation associations 3.3.0 Open spaces with little or no vegetation 3.3.5 Glaciers and perpetual snow	3.1 Mixed forest, Broadleaf forest, Coniferous forest 3.2 Grass, Shrub 3.3.0 Bare land 3.3.5 Permanent ice or snow
4 Wetlands	4 Wetlands
5 Water bodies	5 Water

Table 1 Reclassification based on Corine Legend (First Reclassification method)

As depicted in the workflow, once all the steps are achieved and two data set are re-classified based on the same thematic classification, comparison can be performed. Finally, accuracy measurement indicator error matrix and its derived statistics are calculated and analyzed

3. Results and Discussions

3.1 Ontario land cover map (Reference data set) data processing

Four different tiles as seen in the table of contents of the ArcMap window in the image below, covering the Ontario territotry were obtained for carrying out the accuracy asessment.

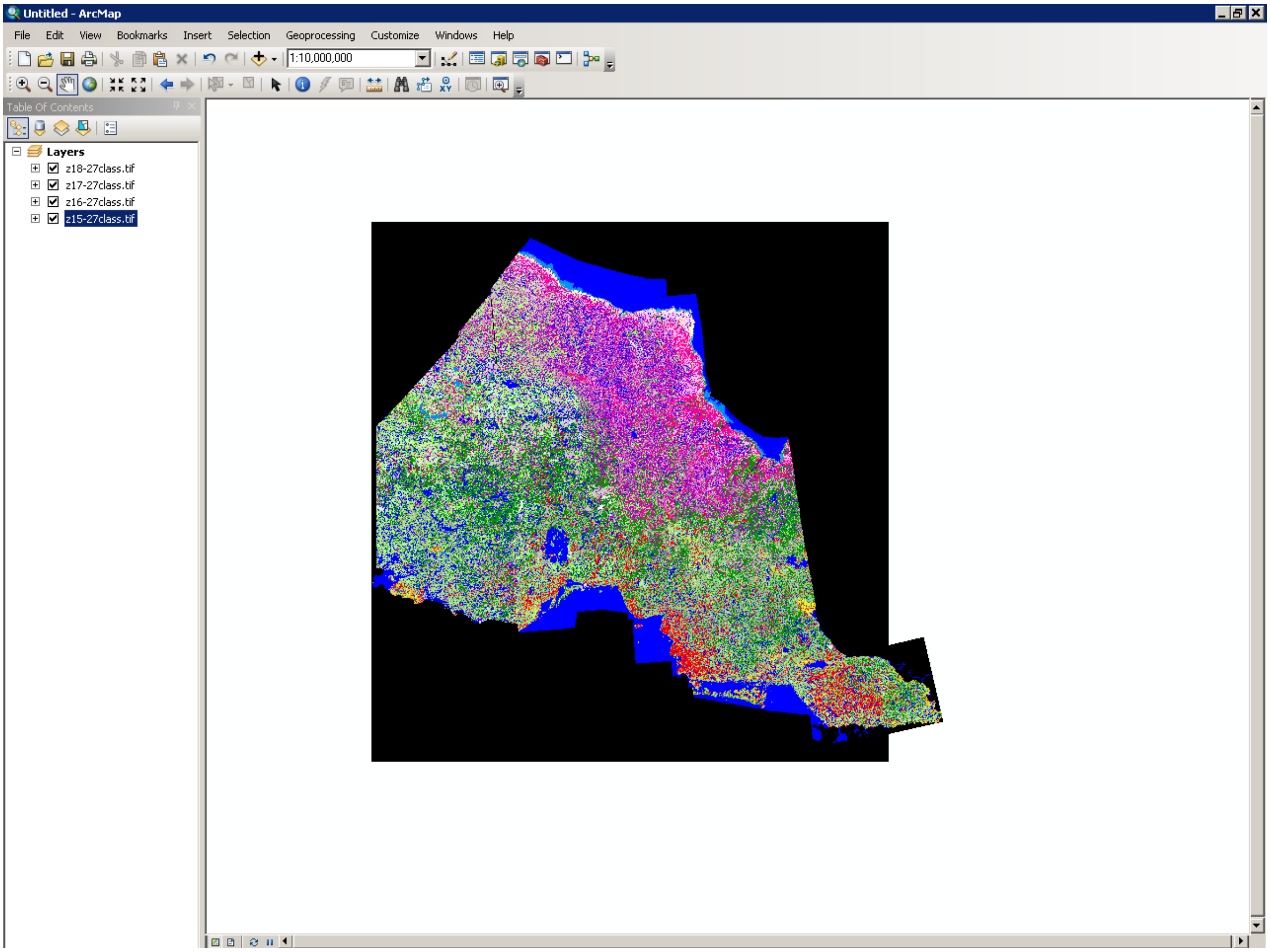


Figure 12 Ontario Region depicted by four tiles

MOSAIC: In order to merge all the four tiles into one, spatial analysis tool named 'Raster Mosaic' has been used on the data set. Figure 13 illustrates the same:

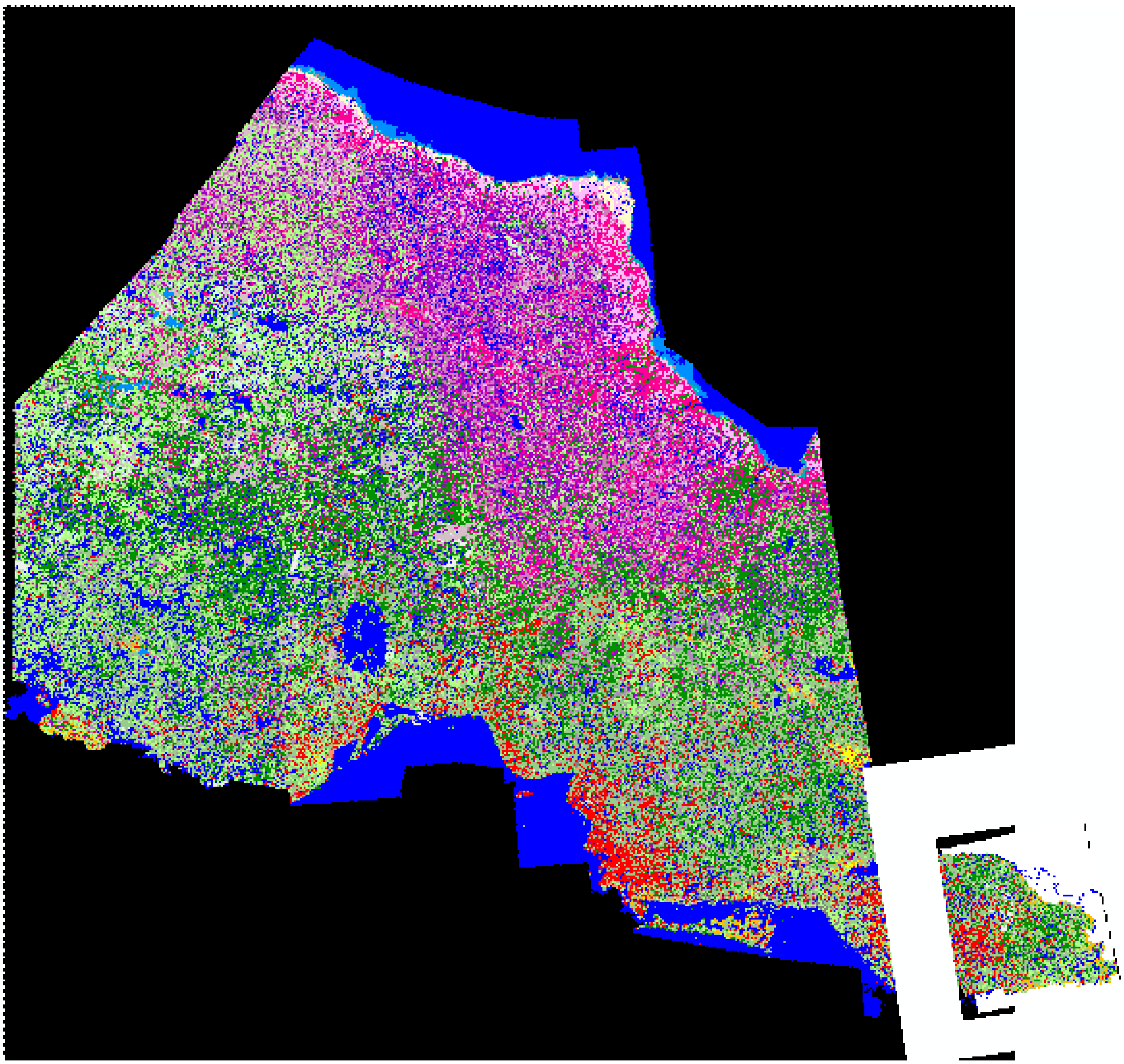


Figure 13 Mosaic raster for the Ontario territory

The attribute table below shows all the land cover classes on the basis of which Ontario land cover map is classfied. The class number is represented by the column ‘Value’ in the attribute table. From the picture above we can see the black area in the image and that black area has been classified as Shadow and therefore the value assigned to it by the software is ‘0’

Table

OntarioLC_final

OBJECTID *	Value	Count	Classes
1	0	53654138	Shadow
2	1	176212058	Water- Deep Clear
3	2	2442006	Water-shallow/sedimented
4	3	2454395	Settlement/Infrastructure
5	4	1130335	Sand/Gravel/Mine Tailings
6	5	5015682	Bedrock
7	6	331915	Mudflats
8	7	27735857	Forest Depletion-cuts
9	8	28604657	Forest Depletion-burns
10	9	16342762	Forest-regenerating depletion
11	10	174269743	Forest-sparse
12	11	60473807	Forest-dense deciduous
13	12	190135607	Forest-dense mixed
14	13	200111073	Forest-dense coniferous
15	15	300204	Marsh-intertidal
16	16	1362031	Marsh-supertidal
17	18	1036	Swamp-deciduous
18	19	5486	Swamp-coniferous
19	20	38108486	Fen-open
20	21	78329376	Fen-treed
21	22	85917055	Bog-open
22	23	109528068	Bog-treed
23	24	4008404	Tundra Heath
24	25	5568491	Agriculture-Pasture/abandoned fields
25	27	976668	Agriculture-cropland
26	28	3106156	Other-unknown
27	29	2776728	Other-cloud/shadow

1 (0 out of 27 Selected)

OntarioLC_final

Table 2 Attribute table for Ontario land cover data based on 27 classes

Class 28: Other-unknown, includes “undefined clearings in disturbed area; small, unburned areas within recent burns; and undefined transitional areas between classes, such as wetland boundaries”

Class29: Cloud and Shadow, includes “areas of cloud or shadow on the satellite images”

RE-PROJECTION: Re-projecting the raster (mosaic) to WGS84 reference system:

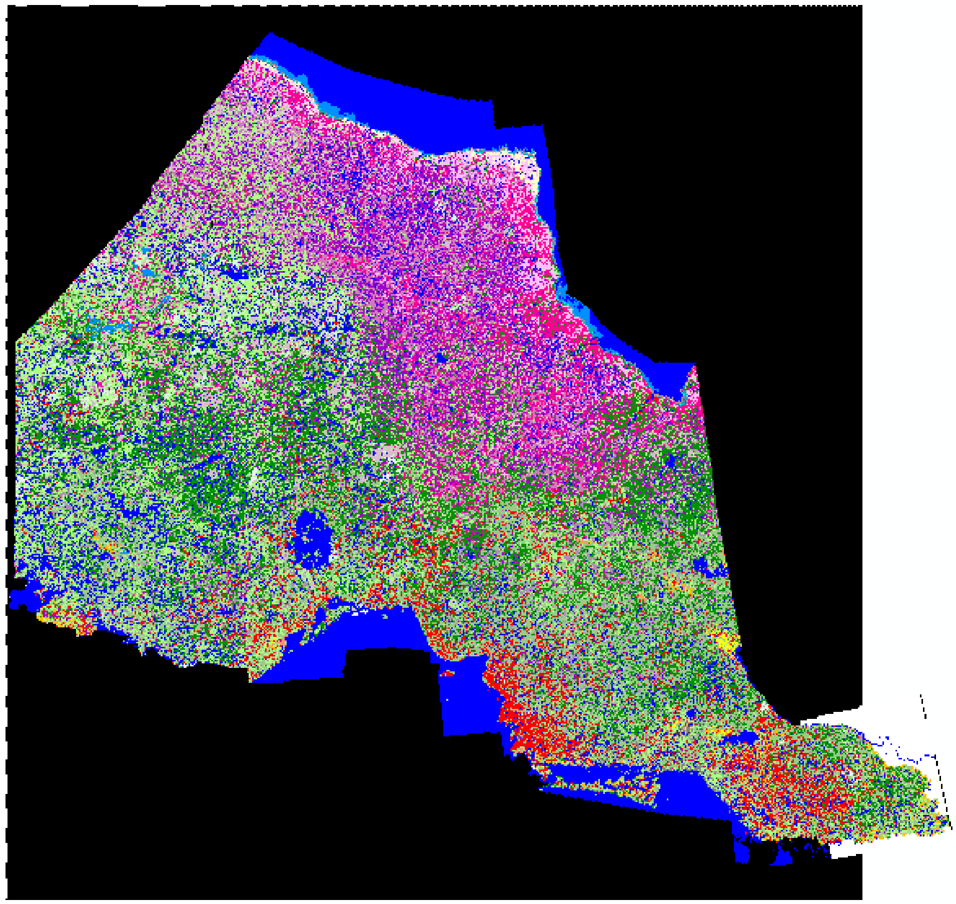


Figure 14 Re-Projected raster for the Ontario territory

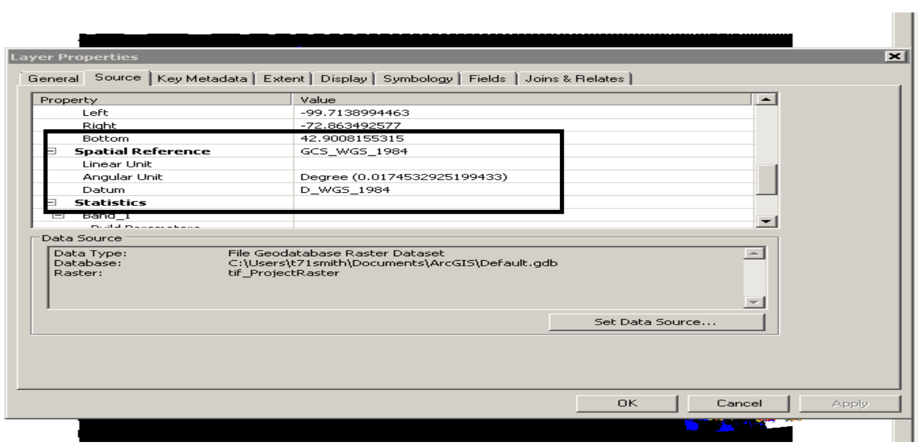


Table 3 Spatial Reference of the re-projected raster (WGS84)

CLIP: Clipping the re-projected Ontario land cover data set with respect to the Ontario boundary shape file and the resulting image is illustrated below in Figure 15.

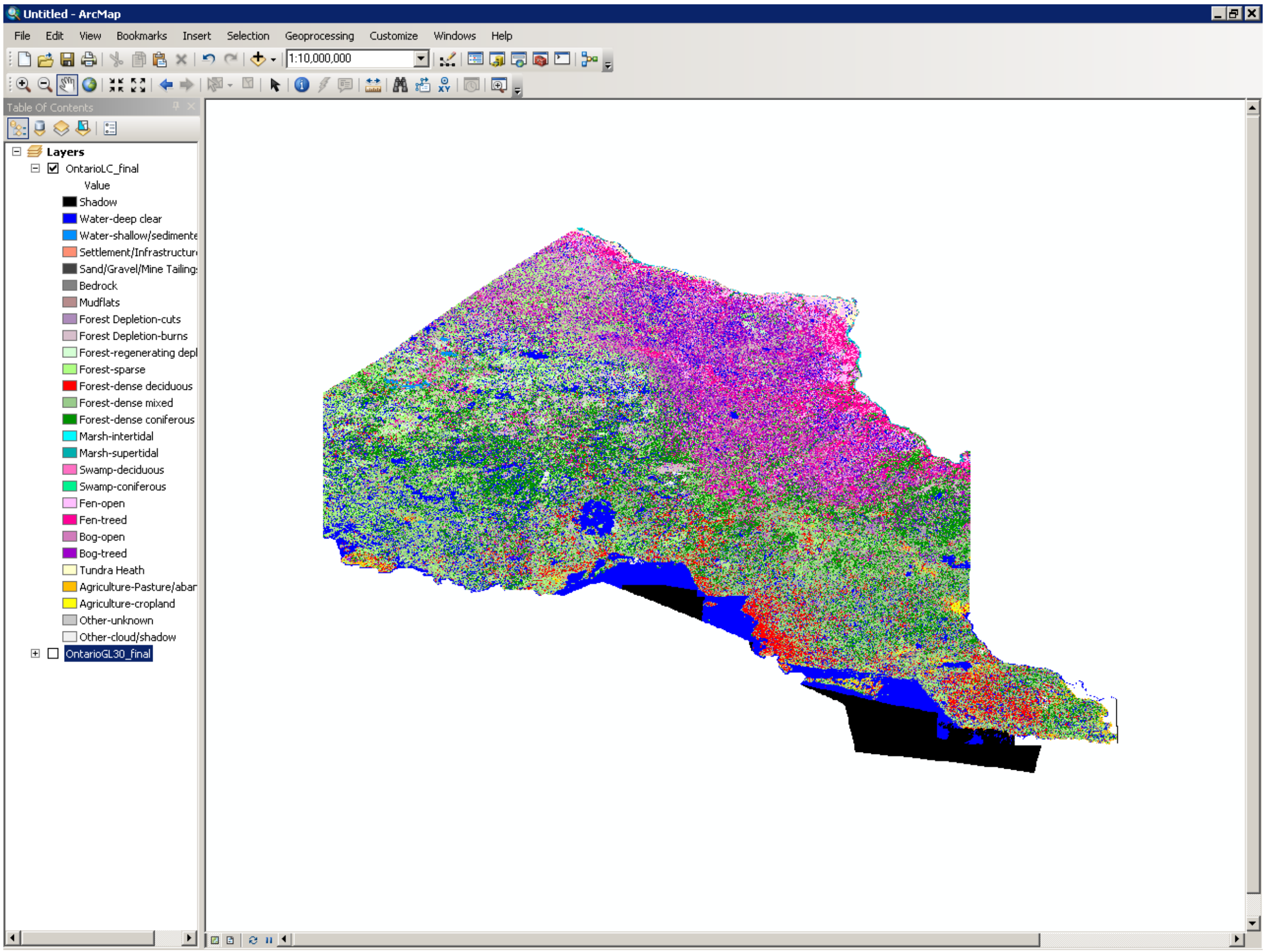


Figure 15 Clipped raster with respect to Ontario polygon boundary layer

FINAL PRODUCT: Ontario land coverage map used in this study.

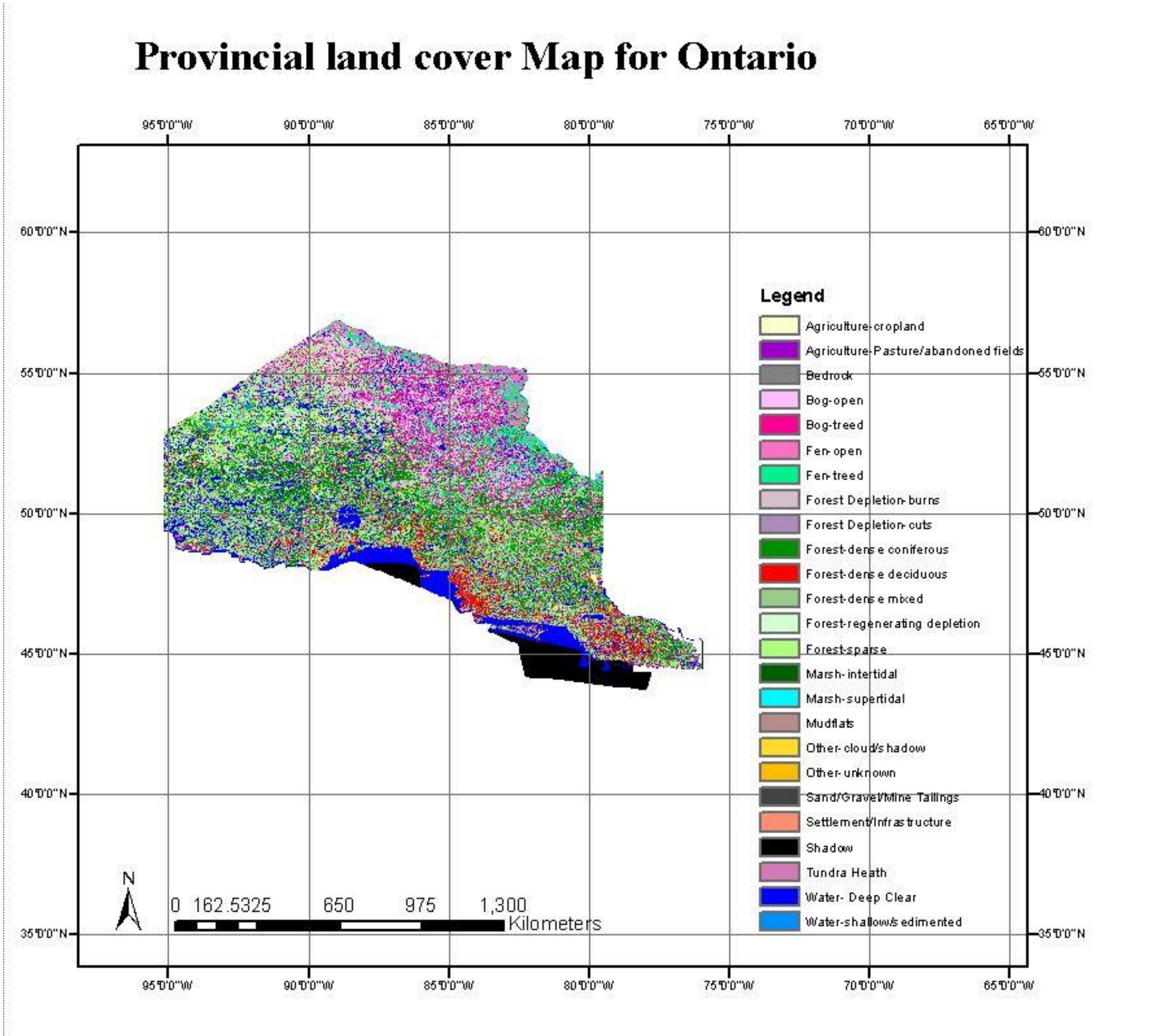


Figure 16 Ontario Provincial Land Cover Map

3.2 GlobeLand30 data processing

The image below in Figure 17 illustrates the GlobeLand30 data set available in the form of 12 different tiles covering the Ontario territory:

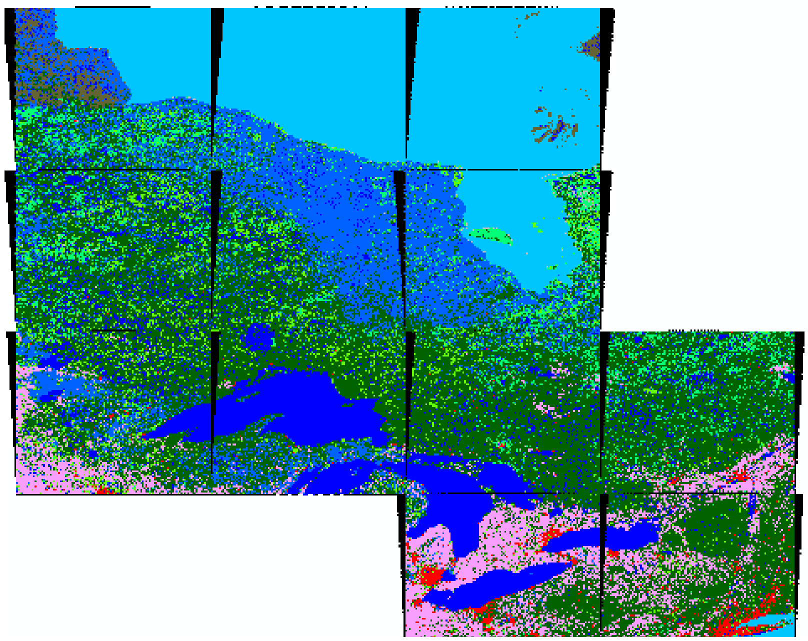


Figure 17 GlobeLand30 data set in the form of 12 tiles

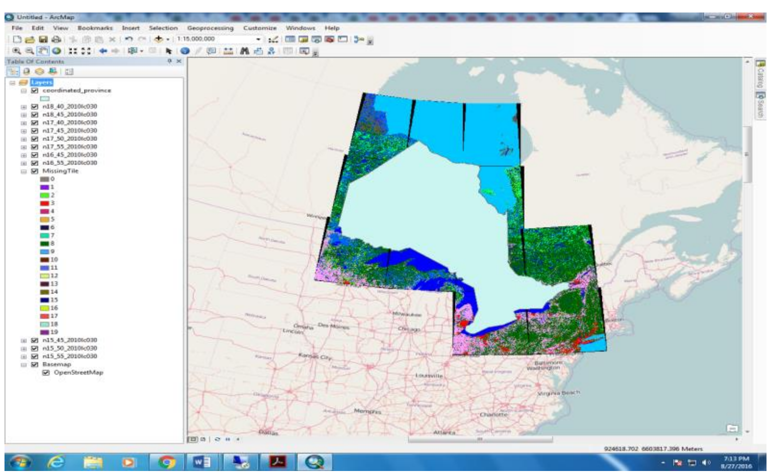


Figure 18 GlobeLand30 data set and Ontario provincial boundary is overlay on base map

MOSAIC and CLIP: Mosaic and Clipping GlobeLand30 data set with respect to the re-projected Ontario province boundary shape file

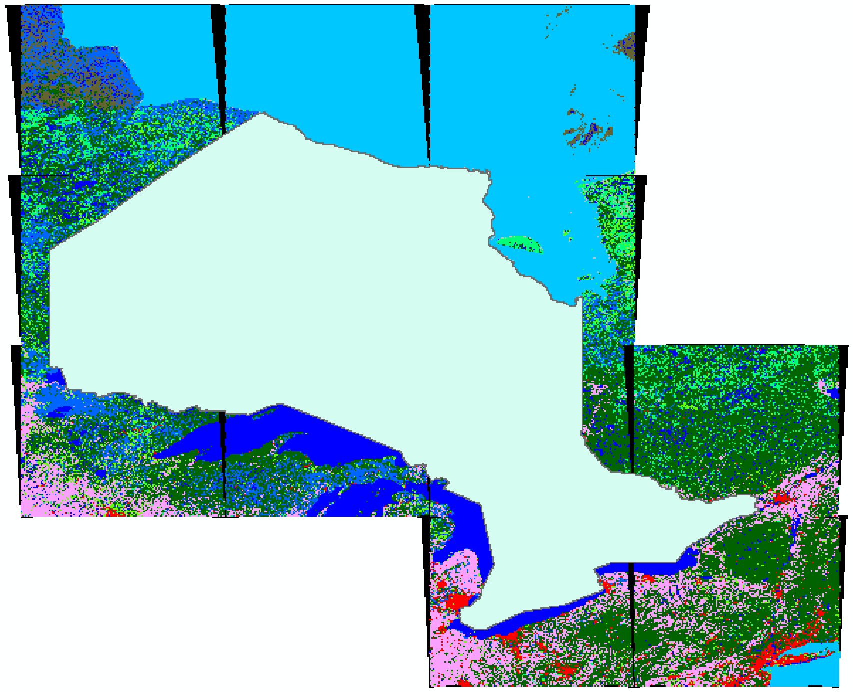


Figure 19 Clipping GlobeLand30 with respect to Ontario polygon boundary layer

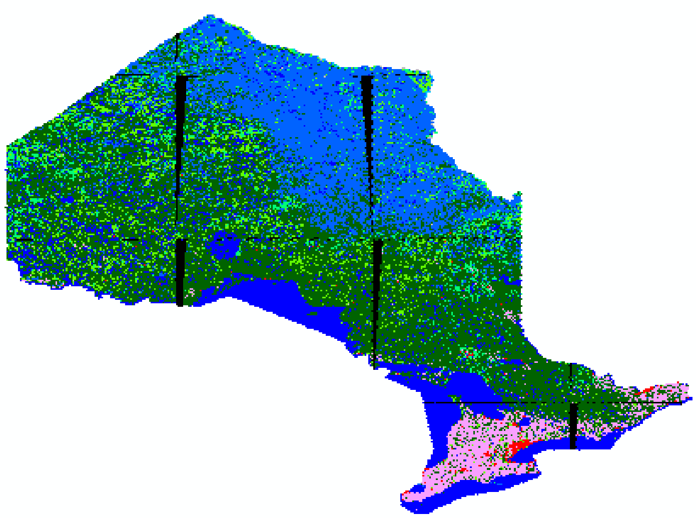


Figure 20 Clipped GlobeLand30 final product

FINAL PRODUCT: GlobeLand30 data for Ontario territory Map

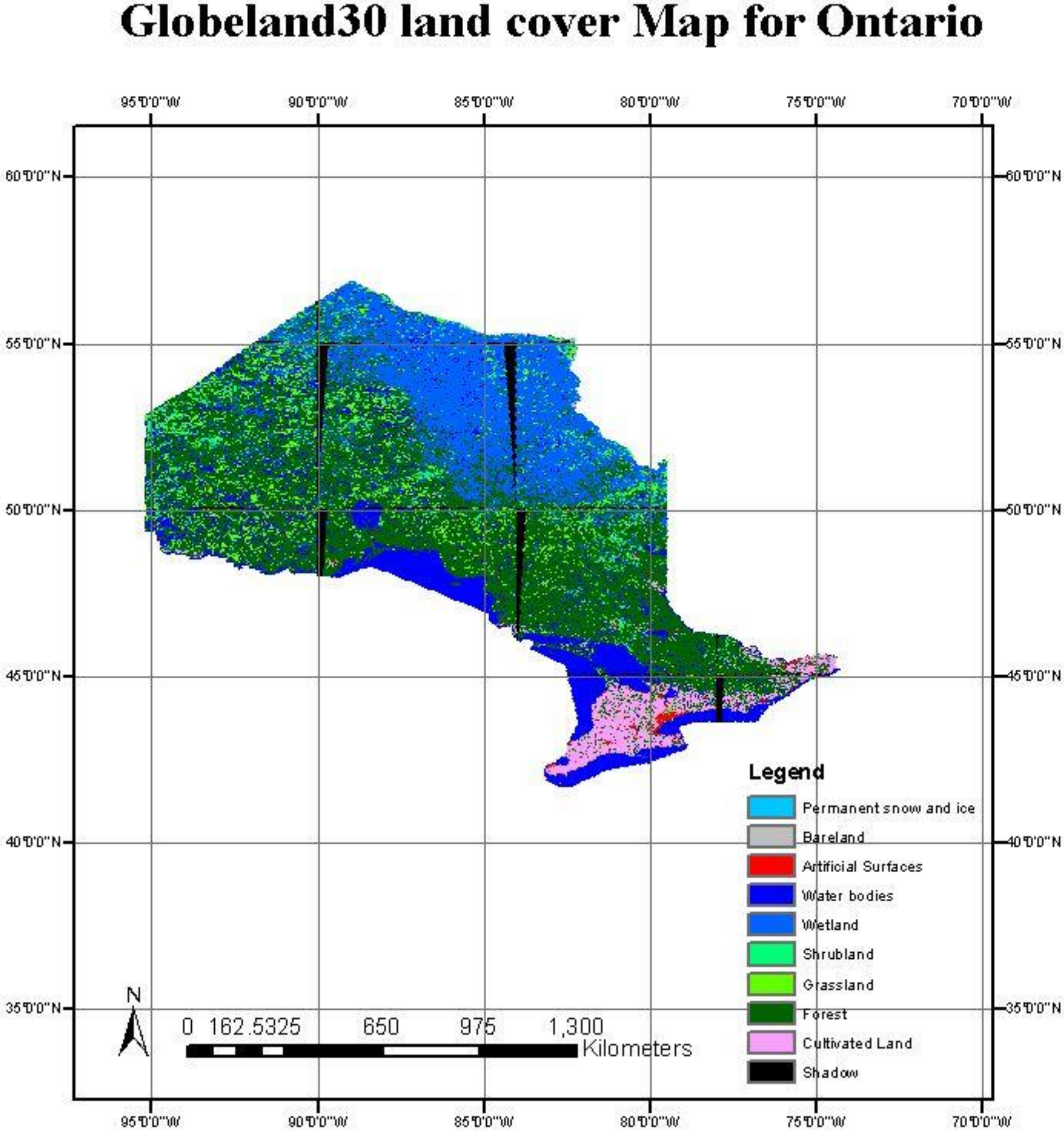


Figure 21 GlobeLand30 Land Cover Map

3.3 Reclassification

3.3.1 GlobeLand30

Table 4 below illustrates how reclassification is done by assigning new code/type to the land cover classes based on the first level of CLC nomenclature (Maria et al. 2015). In the current case study, reclassification method followed takes first five levels of CLC nomenclature for reclassification (Maria et al. 2015) shown in Table 4. This method has been considered over the second reclassification method used in Maria et al. 2015 as it gives the overall accuracy between 81% and 92% while the other approach based on Corine subclasses gave overall accuracies ranging between 62% and 81% (Maria et al. 2015).

(*) Land cover type Tundra (70) in the original classification system is not available for the study area.

Original type/Code*	Translated type/Code
Cultivated land (10)	Agricultural areas (2)
Forest (20)	Forest and semi natural areas (3)
Grassland (30)	Forest and semi natural areas (3)
Shrub land (40)	Forest and semi natural areas (3)
Wetland (50)	Wetlands (4)
Water Body (60)	Water Bodies (5)
Artificial surfaces (80)	Artificial surfaces (1)
Bareland (90)	Artificial surfaces (1)
Permanent snow and ice (100)	Artificial surfaces (1)

Table 4 Table showing new values for GlobeLand30 class merge

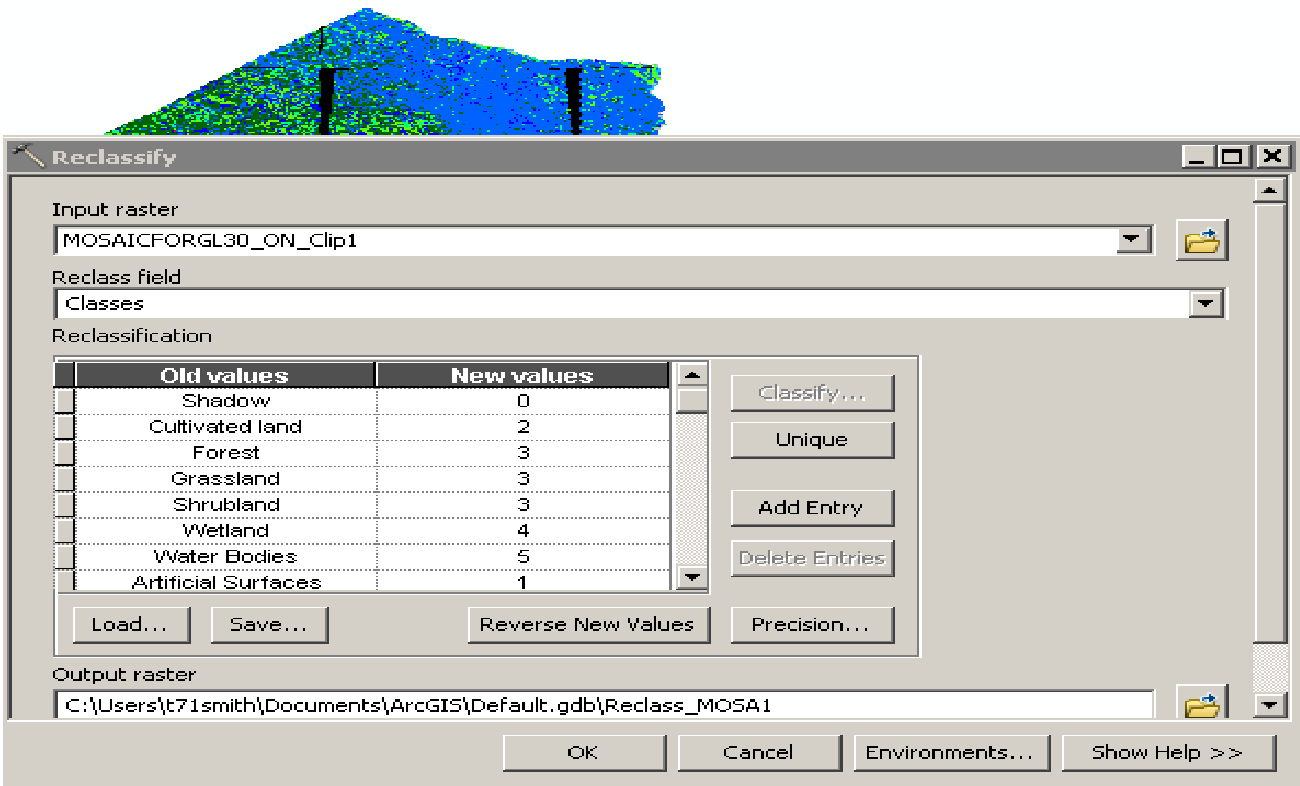


Table 5 Reclassify Tool for reclassifying the data based on CLC nomenclature

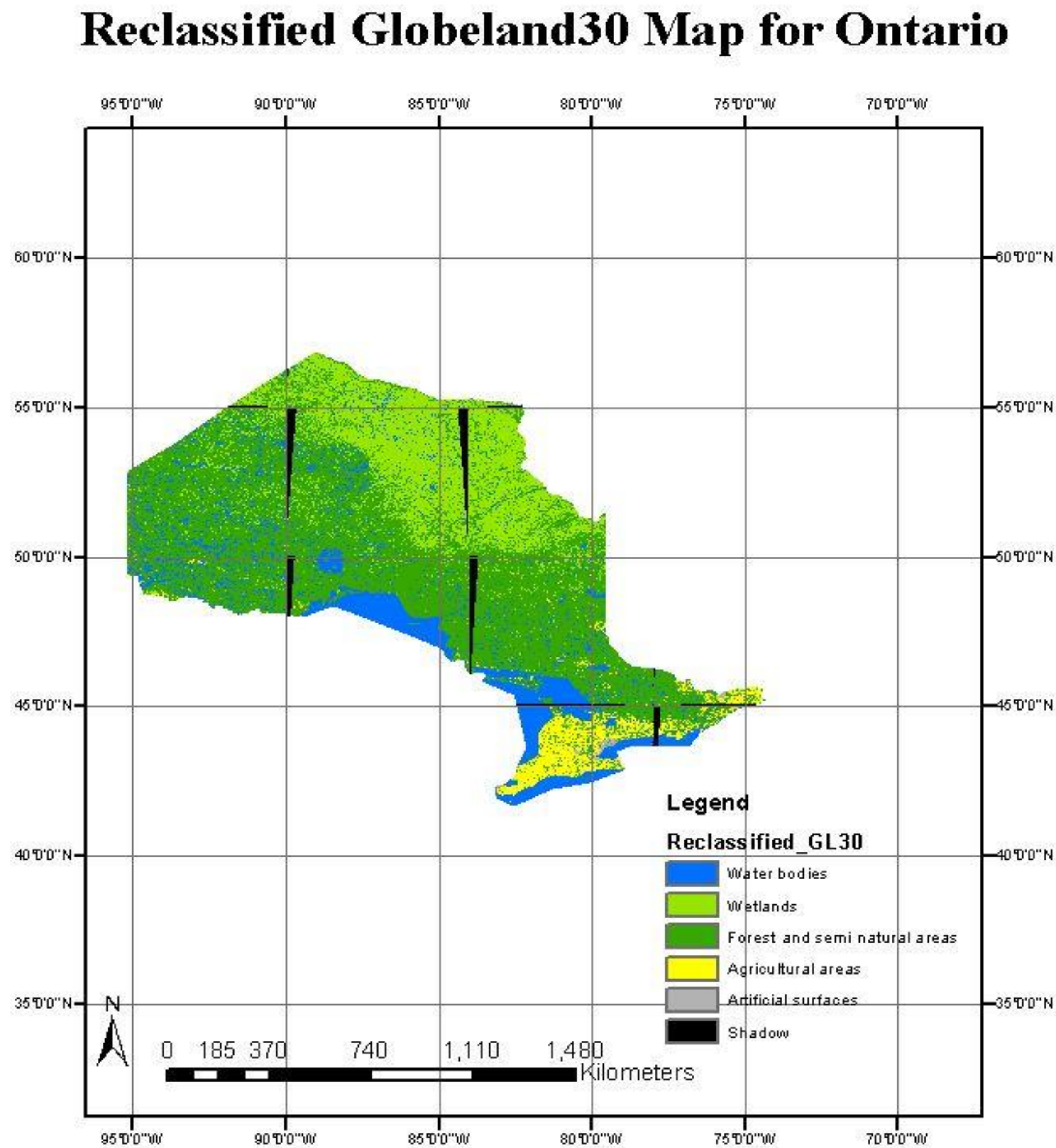


Figure 22 Reclassified GlobeLand30 Map

3.3.2 Ontario data set (Reference)

Table 6 illustrates the new type/code for Ontario data sets based on first five levels of CLC nomenclature reclassification technique (Maria et al. 2015).

Original type/Code	Translated type/Code
Water – Deep or Clear (1)	Water Bodies (5)
Water – Shallow or Sedimented (2)	Water Bodies (5)
Settlement/Infrastructure (3)	Artificial surfaces (1)
Sand/Gravel/Mine Tailings(4)	Forest and semi natural areas (3)
Bedrock (5)	Forest and semi natural areas (3)
Mudflats (6)	Wetlands (4)
Cutovers (7)	Forest and semi natural areas (3)
Burns (8)	Forest and semi natural areas (3)
Regenerating Depletion (9)	Forest and semi natural areas (3)
Sparse Forest (10)	Forest and semi natural areas (3)
Deciduous Forest(11)	Forest and semi natural areas (3)
Mixed Forest (12)	Forest and semi natural areas (3)
Coniferous Forest (13)	Forest and semi natural areas (3)
Intertidal Marsh (14)	Wetlands (4)
Supertidal Marsh (15)	Wetlands (4)
Inland Marsh (16)	Wetlands (4)
Deciduous Swamp (17)	Wetlands (4)
Coniferous Swamp (18)	Wetlands (4)
Open Fen (19)	Wetlands (4)
Treed Fen (20)	Wetlands (4)
Open Bog (22)	Wetlands (4)
Treed Bog (23)	Wetlands (4)
Tundra Heath (24)	Forest and semi natural areas (3)
Pasture (25)	Forest and semi natural areas (3)
Cropland (27)	Agricultural areas (2)

Table 6 Table showing new values for Ontario data sets

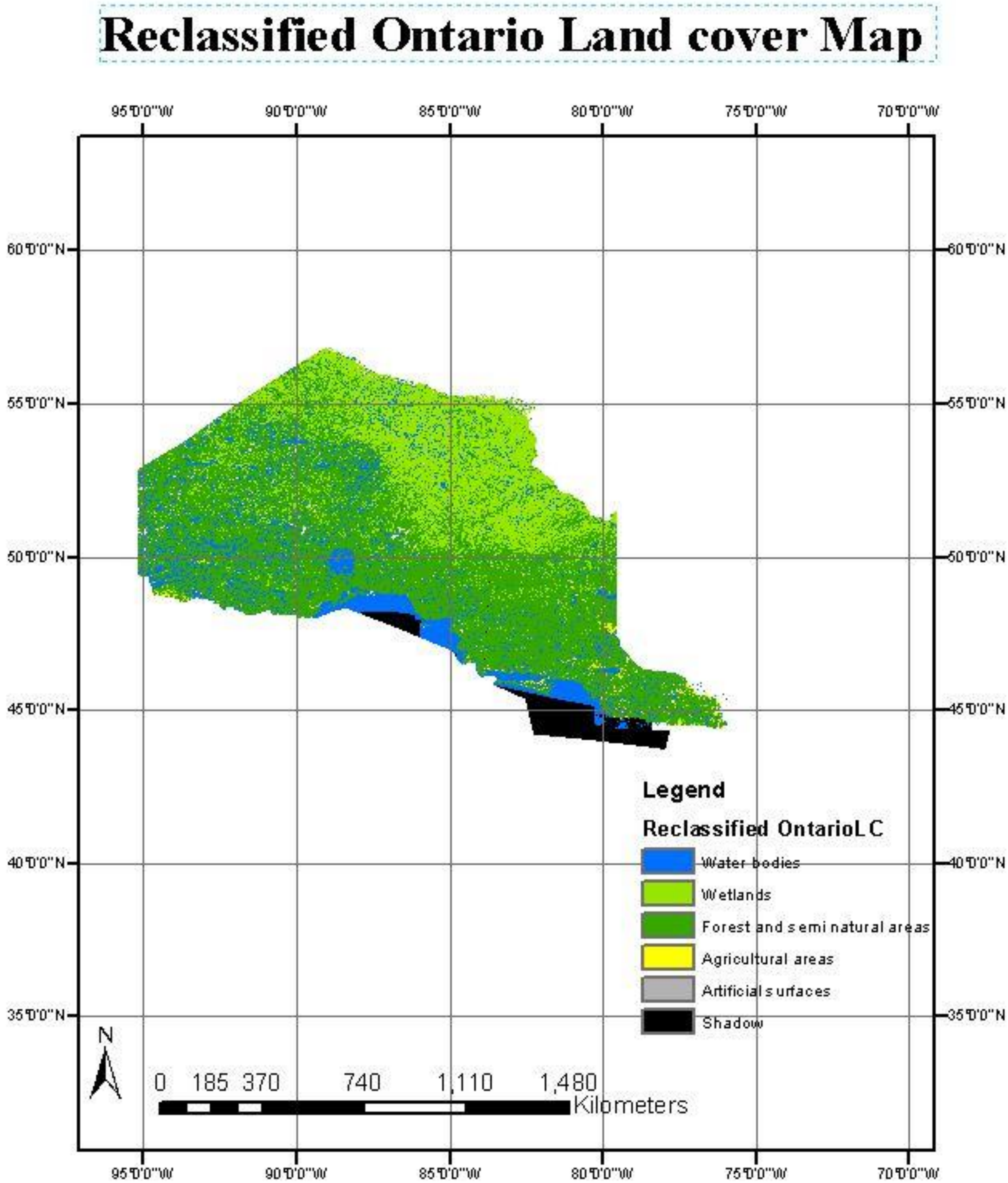


Figure 23 Reclassified Ontario Map

3.4 Accuracy assessment

After classifying, the data sets have the same thematic classification and hence, comparison can be drawn between the two data sets. For accuracy assessment, another image is produced taking into consideration both the reclassified images using equation (7) in Raster Calculator. The reason behind creating this new image is to make comparison between two images (reference and classified) easier as the attribute table we get for the new image after applying equation (7) give values as shown in Table 7 thereby, making it easy to interpret and calculate error matrix. Equation 7 derived from of my own understanding of comparison between two raster images.

$$\textbf{[Reference Image] * 10 + [Classified Image]} \qquad (7)$$

Where,

Reference Image refers to the Ontario land cover data set and Classified Image refers to GlobeLand30 data set for Ontario.

Once this expression is applied the resulting image obtained is a gray scale image as shown below:

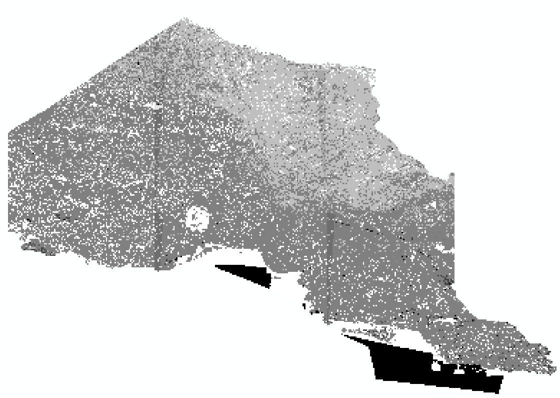


Figure 24 New calculated raster based on equation 7

The attribute table for the resulting image is illustarted below and the pixel value explaination is as follows:

Table

rastercalc1

OBJECTID *	Value	Count
1	0	1371852
2	1	810852
3	2	12443226
4	3	7141097
5	4	291431
6	5	24090196
7	10	45573
8	11	674146
9	12	162222
10	13	1199598
11	14	9398
12	15	22026
13	20	71175
14	21	84130
15	22	3554090
16	23	1886667
17	24	9245
18	25	24135
19	30	19729819
20	31	823386
21	32	1717217
22	33	554301103
23	34	25151091
24	35	3937360
25	40	8835156
26	41	25034
27	42	44385
28	43	30657868
29	44	229264368
30	45	874337
31	50	3224251
32	51	30921
33	52	22886
34	53	21827353
35	54	6545180
36	55	122017849

Table 7 New raster's attribute table

For instance:

<u>(Reference Image)*10 + Classified image = Pixel value in the new calculated raster image.</u> Pixel value 13 for ObjectID 10 means that pixel is classified as "Class 1" in Reference image (Ontario land cover data set) and "Class 3" in Classified image (GlobeLand30 data set). Similarly for pixel value 5 for ObjectID 6, it means that pixel is classified as "Class 0" in Ontario land cover

data set and "Class 5" in GlobeLand30 data set. In the same way we can interpret the rest of the pixel values and by summarizing the number of pixel (Pixel Count) in each class, entries of Error Matrix can be populated. Error Matrix for this study is calculated based on both Pixel count as well as Percentage. The result indicates an overall accuracy of 84% (equation 2) with kappa coefficient of 0.979 (equation 1) for the study area.

Error Matrix: Error matrix is calcluated based on pixel count

		GL30						Total	Producer's Accuracy
		Class 0	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**		
ON LC	**Class 0**	1371852	810852	12443226	7141097	291431	24090196	46148654	2.972680417
	Class 1	45573	674146	162222	1199598	9395	22026	2112960	31.90528926
	Class 2	71175	84130	3554090	1886667	9245	24135	5629442	63.13396603
	Class 3	19729819	823386	1717217	554301103	25151091	3937360	605659976	91.52018046
	Class 4	8835156	25034	44385	30657868	229264368	874337	269701148	85.00681947
	Class 5	3224251	30921	22886	21827353	6545180	122017849	153668440	79.40332381
Total		33277826	2448469	17944026	617013686	261270710	150965903	1082920620	
User's Accuracy		4.1224207	27.53337	19.80654	89.836112	87.74974	80.824773		

Table 8 Error Matrix and derivatives: Producer Accuracy and User Accuracy

where the abbreviations and different classes represent the following: ON LC is Ontario Land Cover,GL30 is Globeland30 data set, Class 0,1,2,3,4,5 are Shadow, Artificial Surfaces, Agricultural areas, Forests and semi natural areas, Wetlands, Water bodies

Total Accuracy = 84.1419% and **Kappa Coefficient**= 0.979

- Forests and semi natural areas (Class 3) shows the highest producer as well as user's accuracy. For shadow (Class 0) user's and producer accuracy are comparatively the lowest of all the classes which is a good sign as it shows that majority of the study area is classfied with respect to the rest of the land cover classes and a very small portion of the GlobeLand30 is classified as Shadow.

- Artificial Surfaces (Class 1) has a producer accuracy of 31.9 % while a user accuracy of 27.5% which depicts that 31.9% of the artificial surfaces have been correctly identified as artificial area however, 27.5% of it is actually artificial surfaces which is not bad at all. Thus, we see that if 31.9% of the times the producer (GlobeLand30) of this map can claim that 31.9% of the time an area that is artificial is identified, user of this map will find 27.5% of the times an area he visits that the map denotes artificial surfaces will actually be artificial surface.
- Also if we look closely into the error Matrix we can see that the producer and user accuracy for all the classes do not have a huge gap and they all are very closely related. In other words, if an area is classfied as Wetlands then there are very high chances of that area actually being Wetlands.
- In case of Forests and semi natural areas we can see the producer accuracy is 91.5% and user accuarcy is 89.83% which again provides a very good relationship between the producer (GlobeLand30) and the user (anyone in the World). This statistic shows that if 91.5% of an area is correctly identified as Forest and semi natural areas by the producer then 89.83% of the area called Forest and semi natural area is actually that which is a great analysis. It is illustrated in Figure 25 and Figure 26.
- Coming to agricultural area, it shows that 63% of an agricultural area has been correctly identified as agricultural area, only 19.8% of the area called agricultural area are actually agricultural area.
- Water bodies by far shows the best results as compared to rest of the classes because the producer and the user's accuracy is in almost the same which means that when 79.4% of an area is correctly identified as water bodies, then 80.82% of the times the area would actually be water body.
- Total accuracy is 84% which explains that the agreement between Classified data set (GlobeLand30) and the reference data set (Ontario) is great thereby, we can conclude that GlobeLand30 data set is an accurate and reliable data source for land cover information.

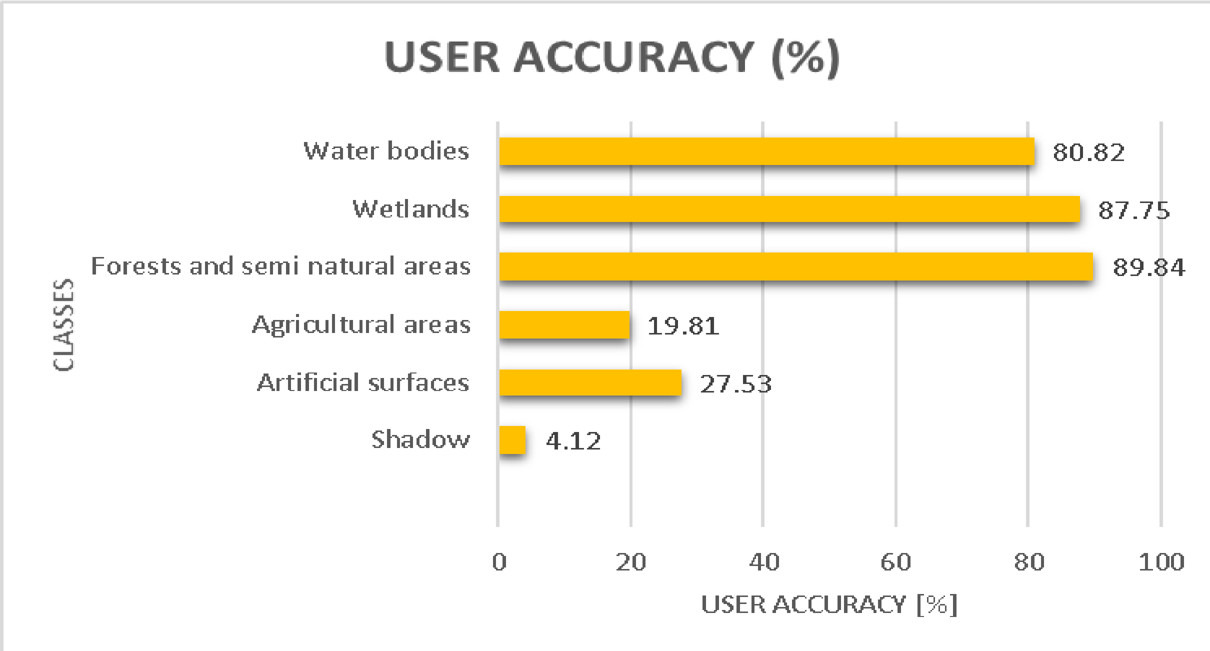


Figure 25 User Accuracy Graph

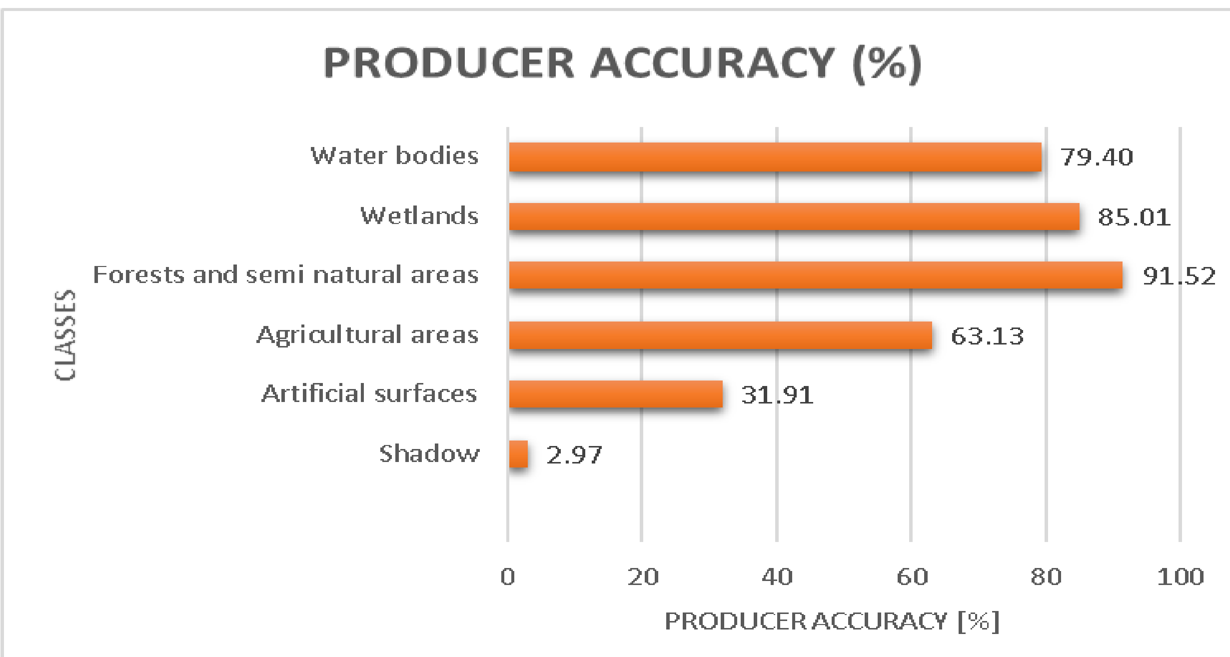


Figure 26 Producer Accuracy Graph

4. Conclusion

In the current case study, the validation of Globeland30 data sets using Ontario data resulted in overall accuracy of 84%, thus proves GlobeLand30 to be one of the most reliable sources considered for future use. The reclassification method selected is the one which is based on five Corine Land Cover (CLC) classes as the overall accuracy obtained is between 81% and 92% for both the baseline years 2000 and 2010 (Maria et al. 2015). In addition, the fact is justified after calculating the overall accuracy for this study and the value is 84% that lies within the expected range of the methodology followed (Maria et al. 2015). One of the reasons for minor disagreement in classification between GlobeLand30 and Ontario regional land cover maps can be due to the images taken in different period and time zones. Secondly, both the maps classified based on different classification system originally. Reclassification for both the maps has been done manually by assigning same legends based on CLC nomenclature (Maria et al. 2015). As it is done manually, the person performing reclassification could introduce errors and discrepancy. The errors could have also been introduced by original developers. However, overall accuracy could largely be improved if the amount of disagreement between both the maps are eliminated. One of the methods to achieve this is by using buffer around the class polygon borders.

In this study particularly it is noticed that while performing mosaic and re-projection method some amount of information is lost and thereby resulting in lower user and producer accuracy for each land cover classification in both data sets. Another important point of discussion is connected to the reference data. Although for carrying out comparison with the GlobeLand30 data set, Ontario regional land cover map is supposed to be the accurate representation of ground reality. The classification method used to classify Ontario data sets is different as compared to GlobeLand30 data sets, therefore, errors pertaining to that is also included in the data sets. Thus, these errors are included in the benchmarking process and can cause anomalies in the validation results (Foody et al. 2010; Foody et al. 2009). Suggestions for future work include analysis on the classification quality of reference data. Practices such as traditional monitoring done in situ and popular approaches based on people’s view will be undertaken (Comber et al. 2013). Additional enhancement measures for this study apart from the methods used to assess the accuracy can also include analysis of correlation and co-variance of errors associated with each class. In addition, detailed and more accurate Ontario data sets are recommended for this study. The problems experienced summarized below:

- Heterogeneous classification methods are followed;
- Number and type of classes is different; and
- Proper classification of land cover classes is not obvious because of the problems related to the data acquisition. In this case, satellites for image acquisition is used and thus radiometric, geometric corrections are common, and the period of the year when the images were captures, presence of cloud that makes the resolution way lower as compared to the images without cloud.

In order to address the issues it is recommended to create a platform such as open global land coverage geo-platform where scientists and other researchers from all around the world can share their views, data and queries thereby, improving the quality of land cover classification. Web-based services like ArcGIS online or any other open geo-portal to share and compare land cover maps to analyze the similarities and highlight the differences between the uploaded shared maps from other sources. Emerging new technologies can also assist in solving such difference with the help of mobile platforms. Examples of such platforms are namely VIEW-IT (Virtual Interpretation of earth Web-Interface Tool) (Clark et al. 2011) and Geo-Wiki project (Fritz et al. 2009).

With the advent of Digital Revolution and advancement and invention of sophisticated spatial data analytical tools, along with the availability of more reliable and consistent Global Land Cover (GLC) data set, the future of Global Land Cover mapping will bring about a revolutionary change to make the world a better habitat. GlobeLand30 is one such initiatives.

References

Anderson, J. R. (1976). *A land use and land cover classification system for use with remote sensor data* (Vol. 964). US Government Printing Office.

Arsanjani, J. J., & Vaz, E. (2015). An assessment of a collaborative mapping approach for exploring land use patterns for several European metropolises. *International Journal of Applied Earth Observation and Geoinformation*, *35*, 329-337.

Arsanjani, J. J., Helbich, M., & de Noronha Vaz, E. (2013). Spatiotemporal simulation of urban growth patterns using agent-based modeling: the case of Tehran. *Cities*, *32*, 33-42.

Arsanjani, J. J., Helbich, M., & Mousivand, A. J. (2014). A morphological approach to predicting urban expansion. *Transactions in GIS*, *18*(2), 219-233.

Arsanjani, J. J., Mooney, P., Helbich, M., & Zipf, A. (2015). An exploration of future patterns of the contributions to OpenStreetMap and development of a Contribution Index. *Transactions in GIS*, *19*(6), 896-914.

Arsanjani, J. J., Mooney, P., Zipf, A., & Schauss, A. (2015). Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. In *OpenStreetMap in GIScience* (pp. 37-58). Springer International Publishing.

Arsanjani, J. J., Zipf, A., Mooney, P., & Helbich, M. (2015). An introduction to OpenStreetMap in Geographic Information Science: Experiences, research, and applications. In *OpenStreetMap in GIScience* (pp. 1-15). Springer International Publishing.

Arsanjani, J. J., Zipf, A., Mooney, P., & Helbich, M. (2015). An introduction to OpenStreetMap in Geographic Information Science: Experiences, research, and applications. In *OpenStreetMap in GIScience* (pp. 1-15). Springer International Publishing.

Ban, Yifang. (2014). GlobeLand30: A Milestone in Earth Observation and Open Access Geospatial Information. GIM International, Mapping the World. N.p., Web.

Biagi, L., & Negretti, M. A new approach to the polygons rasterization in GRASS.

Brovelli, M. A., Molinari, M. E., Hussein, E., Chen, J., & Li, R. (2015). The first comprehensive accuracy assessment of GlobeLand30 at a national level: methodology and results. *Remote Sensing*, *7*(4), 4191-4212.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., ... & Zhang, W. (2015). Global land cover mapping at 30m resolution: A POK-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, *103*, 7-27.

Chen, Jun, Ban Y, and Li S. (2015). "China: Open Access to Earth Land-cover Map." *ISPRS Journal of Photogrammetry and Remote Sensing* n. pag. Web.

Chen, Lijun. (2014). "GlobeLand30 Product Description." National Geomatics Center of China (NGCC). Web.

Clark, M. L., & Aide, T. M. (2011). Virtual interpretation of Earth Web-Interface Tool (VIEW-IT) for collecting land-use/land-cover reference data. *Remote Sensing*, *3*(3), 601-620.

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., & Foody, G. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. *International Journal of Applied Earth Observation and Geoinformation*, *23*, 37-48.

Congalton, R. G., & Green, K. (1999). Assessing the accuracy of remotely sensed data: principles and applications. *Lewis Pub—lishers, Boca Raton, Fla*, 137

Eklund, Lina. (2014). Friday Map: Land Cover in the Duhok Governorate According to GlobeLand30. Population and Environment in the Middle East. Web.

ESRI. (2016). "Data Quality and Validation." ArcGIS for Desktop. N.p. Web.

Feddema, J. J., Oleson, K. W., Bonan, G. B., Mearns, L. O., Buja, L. E., Meehl, G. A., & Washington, W. M. (2005). The importance of land-cover change in simulating future climates. *Science*, *310*(5754), 1674-1678.

Foody, G. M. (2009). The impact of imperfect ground reference data on the accuracy of land cover change estimation. *International Journal of Remote Sensing*, *30*(12), 3275-3281.

Foody, G.M. (2011). Classification accuracy assessment. *IEEE Geosci. Remote Sens. Soc. Newsletter,* 159, 8–14.

Jokar Arsanjani, J., Helbich, M., Bakillah, M., Hagenauer, J., & Zipf, A. (2013). Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science*, *27*(12), 2264-2278.

Jokar Arsanjani, J., See, L., & Tayyebi, A. (2016). Assessing the suitability of GlobeLand30 for mapping land cover in Germany. *International Journal of Digital Earth*, 1-19.

Liu, C., Frazier, P., & Kumar, L. (2007). Comparative assessment of the measures of thematic classification accuracy. *Remote sensing of environment*, *107*(4), 606-616.

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, *148*, 42-57.

Olofsson, P., Foody, G. M., Stehman, S. V., & Woodcock, C. E. (2013). Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, *129*, 122-131.

Pontius Jr, R. G., & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*(15), 4407-4429.

Pringle, M. J., Schmidt, M., & Muir, J. S. (2009). Geostatistical interpolation of SLC-off Landsat ETM+ images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *64*(6), 654-664.

Scholars GeoPortal. (2015). "Scholars GeoPortal BrowserTab." N.p. Web.

Spectranalysis Inc., and Remote Sensing And GIS Services. (2004). Introduction To The Ontario Land Cover Data Base. *Spectranalysis*. Print.

Stehman, S. V. (2009). Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing*, *30*(20), 5243-5272.

Stehman, S. V., & Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, *64*(3), 331-344.

Stehman, S. V., & Wickham, J. D. (2011). Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. *Remote Sensing of Environment*, *115*(12), 3044-3055.

Strahler, A., Boschetti, L., Foody, G. M., Fiedl, M. A., Hansen, M. C., Herold, M., ... & Woodcock, C. Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment Of Global Land Cover Maps Report of Committee of Earth Observation Satellites (CEOS)—Working Group on Calibration and Validation (WGCV), 2006. *JRC report series*.

Sun, B., Chen, X., & Zhou, Q. (2016). Uncertainty Assessment of GLOBELAND30 Land Cover Data Set Over Central Asia. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1313-1317.

Thomlinson, J. R., Bolstad, P. V., & Cohen, W. B. (1999). Coordinating methodologies for scaling landcover classifications from site-specific to global: Steps toward validating global map products. *Remote Sensing of Environment*, *70*(1), 16-28.

Vaz, E., & Jokar Arsanjani, J. (2015). Crowdsourced mapping of land use in urban dense environments: An assessment of Toronto. *The Canadian Geographer/Le Géographe canadien*, *59*(2), 246-255.

Verburg, P. H., Neumann, K., & Nol, L. (2011). Challenges in using land use and land cover data for global change studies. *Global Change Biology*, *17*(2), 974-989.

Xu, Hong, Sangeet Mathew, and Mark Harris. (2010) "Managing Imagery and Raster Data in ArcGIS." *ESRI International User Conference.*

Zhao, Y., Gong, P., Yu, L., Hu, L., Li, X., Li, C., ... & Cheng, Q. (2014). Towards a common validation sample set for global land-cover mapping. *International Journal of Remote Sensing*, *35*(13), 4795-4814.