Statistical Analysis of Utilization of Landsat Data in Observation of Small Inland Water Bodies

Miroslav Pasler, Jitka Komarkova, Ivana Cermakova
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic
{miroslav.pasler, jitka.komarkova, ivana.cermakova}@upce.cz

***Abstract*—Spatial data are very important for assuring quality of decision-making. Satellite images are used in many applications as a source of spatial data. Precise knowledge of number of usable (suitable) satellite images suitable for analyses is important. Statistics can be used to predict number of suitable images for a particular time period. The determination which images are suitable enough for analysis is the first step. The next step is to analyze the results of the determination of image suitability, especially with respect to distribution of images in time. This is the main topic of this paper. Statistical analyses methods are used. The main aim is to predict how many suitable Landsat images can be expected in the case of small inland water bodies observation. A part of Pardubice region in the Czech Republic is used within the case study to demonstrate the proposed procedure. The proposed procedure can be applied to other areas and regions as well.**

***Keywords*—Landsat; imagery; small water body; remote sensing; statistics**

I. INTRODUCTION

Water quality monitoring and measuring is historically very significant branch in the field of satellite based remote sensing. From the very beginning of satellite imaging, it is widely used in scientific research in water quality observation. In fact, it is one of the most used method of data collection. Works of Bukata, Harris and Bruton [2, 3] are the very early works in the field dealing still with Landsat1 (ERTS-1) data. There were many of works during 70's and 80's as stated by Middleton and Marcell [4]. According to analysis of publications, the main increase of studies came along with Landsat 5 and later Landsat 7 and 8 launches, e.g. for land use changes detection [8], ice dynamics [9], surface water proportion in inland river basins [10], identification of impervious features [11], vegetative cover loss [12], etc. While new methods including UAV and Lidar are approaching, the satellite based remote sensing of small water bodies still plays significant role in nowadays research. Its role is even more important considering the works dealing with Sentinel satellites [1], [7]. However, according to analysis of keywords, the most popular and the most used system in this branch is still Landsat system with Landsat 8 in the leading position of the current research.

Previous works of authors [5, 6] pointed out that there is a lack of scientific researches focusing on evaluation of suitability of remotely sensed images for small water bodies observation. There is a theoretical research available, which deals mostly with design of models for determination of chosen parameters of water quality. There are also practical applications and case studies available in this field. Today, there is a high need for periodic observation of the areas with small inland water bodies. It means that area of water bodies cannot be covered by clouds or image errors. Clouds and image errors are the most important sources of gaps in time series of satellite imagery. Gaps in the time series can influence observation of an area of interest in time. So, there is a need to calculate in advance how long could be the gap be or how many images will be with the highest probability usable in a specific time period.

The previous works of authors [5, 6] are mostly focused on methodology of an evaluation of suitability of Landsat images, on the determinants of the suitability of images and their influencers. Cloud cover and the black image gaps are identified as the main influencers on the suitability of images in the case of Landsat 7 images. Based on the results and on the nature of observation of small inland water bodies, there is also a difference between influence of these factors to larger water bodies observation and influence of these factors in cases of observation of areas with smaller water bodies, which are differently distributed in an area of interest. This is caused by a possible specific distribution of clouds and the black gaps in the image with respect to distribution of the water bodies themselves. It leads to the fact that the percentage of cloud cover in a whole image is not sufficient enough to determine the suitability of images.

The main aim of the paper is to propose a suitable procedure of calculation of a number of usable satellite images (i.e. with visible small inland water bodies) in advance with taking into account cloud cover and errors in sensing. There is no serious research correlating the cloud cover over an image and usability of the image for small inland water bodies observation. There is no statistical analysis available that allows to determine the time gap between images, probability of having a usable image within the given time period or the most probable number of usable images within a specific time period as well. The proposed procedure will be demonstrated on a case study.

Structure of the paper is as follows: the second chapter describes used data and methods. Next, hypotheses and

research questions are stated. Next chapter describes the proposed procedure. The following chapter provides answers to the research questions and hypotheses. Conclusion follows.

II. Data and Methods

A. Area of Interest and Data

An area of interest is located to the north of the city of Pardubice and it contains several small inland water bodies – ponds and small lakes created by mining of sand. The water bodies are predominantly ponds designed for fish breeding and for outdoor swimming. Total water area is approximately 0.1 km^2 and it is spread over a region with an area of 150 km^2 [5, 6]. The water bodies are shown in Landsat 8 image and in map (see Fig. 1).

Figure 1 Landsat 8 image of the chosen water bodies

Data for analyses presented in this paper consist of Landsat 7 and 8 satellite images of the area of interest. Landsat satellites provide moderate-resolution imagery of the Earth's land surface in several various spectrums.

In the case of the study described by this paper, analyzed data set consists of 215 Landsat 7 and Landsat 8 images recorded from March 29th, 2013 to November 15th, 2015. The time gap between particular images is not constant due to used path/row coordinates, used satellites and their flight offset. The time spacing is neither random. It is periodical set of days between the images lasting from 1 to 8 days (16 days in the case of a missing image in rare cases).

B. Methods

For every image, there is calculated percentage of clear water surface as it was mentioned in the previous work [6]. The percentage is an input for determination if the image is usable or not for small water bodies observation. The minimum percentage of the clear water surface can differ with respect to planned analyses, observed parameters and other factors. In this paper, suitability of images is evaluated from the point of different values of the minimum percentage of the clear water surface. The parameter is called γ further in this paper and its values are as follows:

$$\gamma_a \in 0;1;\ a \in \{5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 75, 80, 85, 90, 95\} \quad (1)$$

The a index represents the minimum required percentage of clear water surface to consider the image as usable. The usability itself (γ_a) is considered as a dichotomic variable where 1 represents usable image and 0 unusable image. It implies that the higher value of α parameter the lower number of usable images.

There are calculated numbers of days between following usable images for different values of index a of parameter γ. The result represents time gap between the usable images and it is called δ_a further in this paper, where index a has the same meaning as in case of γ_a parameter. Parameter δ_a is calculated for every image as well as parameter γ_a so the minimum value is 1 and hypothetic maximum depends on the number of days between the first and the last image in data series.

Except the calculated area of free water surface, γ_a and δ_a parameters, there are used other parameters, which represent metadata for the image. These variables are part of the data set due to their potential correlation to outputs and to each other. The variables, their possible values, their type and their brief description are described by the Table 1.

Common tools of statistical induction are used for the data analysis. The significance value of all tests is set to $\alpha = 0.05$.

Let's assume that the acceptability γ_a is a random variable coming from a set with Bernoulli distribution, where the success of random event is presence of a usable image $\gamma_a = 1$ with probability π. So the point estimation is calculated by (2) and interval estimation is calculated by (3) and (4):

$$\pi = \frac{x}{n} \quad (2)$$

where x is number of usable images and n is number of all images in the set.

TABLE I. Variables of Used Data Set

Variable name	Type	Values	Description
Percentage	numeric, continuous	<0 - 100> [%]	Represents percentage of water surface in the image which is not devaluated by presence of clouds, gaps or other influences
γ_a	dichotomic	{0, 1}	Described above
δ_a	numeric, discrete	<1 – 920> [days]	Described above
Date	date	-	Day, month and year, when the image was taken

Satellite	nominal	{L7, L8}	The satellite taking the image
Path	nominal	{191, 192}	Represents the path of the satellite taking the image
Azimuth	numeric, continuous	<0 - 360> degrees	The angle of sun when the image was taken
Elevation	numeric, continuous	<0 - 90> degrees	The elevation of sun when the image was taken
ACCA	numeric, continuous	<0 – 100> [%]	Percentage of clouds over whole image calculated by Automated Cloud Cover Assessment

$$p - \frac{1}{2n} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < \pi < p + \frac{1}{2n} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \quad (3)$$

$$p - \frac{1}{2n} - z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}} < \pi \quad (4)$$

Equations (3) and (4) represent 95 % estimations of parameter π.

Let's assume that parameter n_k is an average number of images taken in a period indexed with k (1 – month, 2 – quarter, 3 – year) so γ_a is a random variable with binomial distribution B(n,π). So, the estimation of this distribution is calculated using the following equations:

$$E(X) = Np \quad (5)$$

for mean, and:

$$D(X) = Np(1-p) \quad (6)$$

for variance.

Equation (7) represents probability that just r images will be usable:

$$P(X = r) = \binom{N}{r} p^r (1-p)^{N-r} \quad (7)$$

So the values of distribution function are estimated as follows:

$$P(X \leq r) = \sum_{k=0}^{r} \binom{N}{k} p^k (1-p)^{N-k} \quad (8)$$

with an assumption of the equation (9).

$$P(X = s) = 1 - P(X \leq r); s = r + 1 \quad (9)$$

III. Hypotheses and Research Questions

The used methods of statistical induction should give an answer to the question described in the introduction and give quantification of the number of usable images (with visible surfaces of small water bodies). The hypotheses are formulated in the following way according to the problem and the described dataset:

1. There is a correlation between values of variables Satellite and Path to variable Percentage
2. There is a correlation between variables ACCA and Percentage
3. There is a significant trend in the variable Percentage considered as time series.

Except for the hypotheses, there are formulated the following research questions to be quantified using the described methods:

1. How long can possibly be the time period without any usable image?
2. What is the probability that in specific period (month, quarter, year) there will be stated number of suitable (usable) images with respect to parameter a?
3. How many images will be suitable (usable) in specific time period with the highest probability?
4. What is the probability of predetermined minimum number of suitable (usable) images in the chosen time period?

IV. Case Study and Data Processing

The area of interest and used data are described in the previous chapter along with the used methods.

The basic descriptive statistics is calculated for the data set. The results for chosen variables are summarized in the Table 2 where summarized number of γ_a, minimum, maximum and average values of δ_a are shown.

The decreasing linear trend of number of usable images according to the parameter a is shown in the Fig. 2 (the left part). There is also shown dependency between a and average time gap between suitable (usable) images (the right part), which is increasing exponentially.

TABLE II. BASIC DESCRIPTIVE STATISTICS (MINIMUM, MAXIMUM AND AVERAGE NUMBER OF DAYS WITHOUT USABLE DATA IS SHOWN)

a		5	10	15	20	25	30	40	50	60	70	75	80	85	90	95
$\gamma_{a\,(sum)}$		85	80	77	71	65	62	60	45	35	29	25	20	20	16	8
δ_a	**Min**	1	1	1	1	1	1	1	1	1	1	1	16	16	16	16
	Max	39	39	39	55	55	55	56	57	105	113	113	113	113	154	273
	Ave	10.8	11.5	11.9	13	14.2	14.8	15.3	20.4	26.3	30.7	35.4	43.8	43.8	54.1	102

	Average	**Minimum**	**Maximum**	**σ**	**N**
Percentage	21.631	**0**	**100**	32.358	**215**
ACCA	58.869	0	99	29.967	**215**

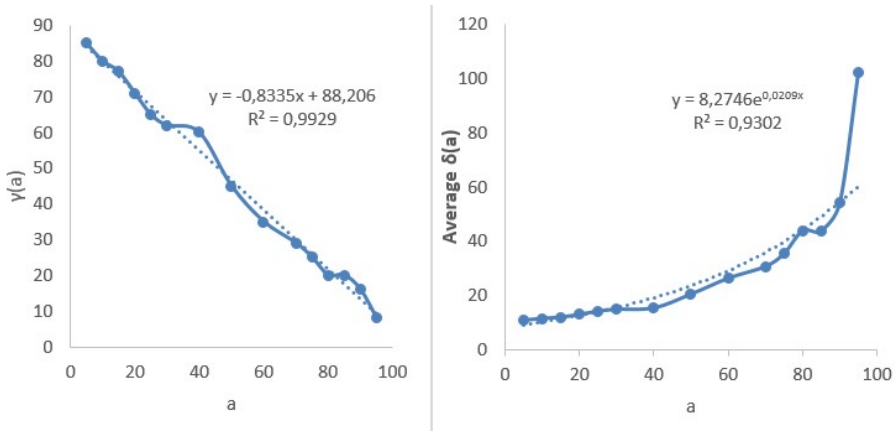

Figure 2 Number of acceptable images and time gap between them dependent on *a* parameter.

Table 3. Parameters of Binomial Distribution of Variable γ_a

a	5	25	50	75	95
			n=7		
E(X)	2.765	2.114	1.463	0.812	0.259
D(X)	1.673	1.475	1.157	0.718	0.249
r	3	2	1	0	0
P(X=r)	0.2889	0.317	0.358	0.422	0.768
s	1	0	0	0	0
P(s=1)	0.971	0.92	0.807	0.578	0.232
			n=21		
E(X)	8.295	6.342	4.389	2.436	0.777
D(X)	5.018	4.427	3.472	2.153	0.748
r	8	6	4	2	0
P(X=r)	0.175	0.187	0.212	0.271	0.453
s	5	3	2	0	0
P(s=1)	0.999	0.999	0.993	0.925	0.547
			n=83		
E(X)	32.785	25.066	17.347	9.628	3.071
D(X)	19.834	17.496	13.721	8.511	2.957
r	33	25	17	9	3
P(X=r)	0.089	0.095	0.107	0.136	0.228
s	26	18	11	5	1
P(s=1)	0.999	0.999	0.999	0.999	0.956

The correlation between variables, namely between ACCA and Percentage, is calculated and the influence of variables Path and Satellite is tested by Kruskal-Wallis test. The correlation coefficient for relation between ACCA and Percentage is -0.769.

There is tested a presence of trend, cyclic and seasonal component in variable Percentage taken as time series. The autocorrelation between values of Percentage is also tested. There is found no autocorrelation as well as no significant component of decomposition of the time series.

According to equations (5) to (9), there are estimated parameters of variable γ_a as a variable with binomial distribution B(n,π). They are calculated for different values of the parameter *a* and for different values of *n*. The parameter *n*

represents an average number of images taken in the time period (7 – month, 21 – quarter, 83 – year). The parameter π is taken from estimation calculated by means of the equation (2). Table 3 shows the results for chosen values of *a*. There is calculated expected value, variance, the most probable number of usable images (*r*), probability of occurrence of *r* (P(X=r)), number of usable images occurred with the probability higher than 0.95 (s) and probability of occurrence of at least one usable image (P(s=1)).

V. Results

Data set analyses provides the following results:

1. Neither variable Satellite, nor variable Path have any significant influence to percentage of clear water surface
2. There is negative relation between the variables. The correlation coefficient is -0.769
3. There is no significant trend in variable Percentage

According to the processing of data, the answers for the questions are following:

1. Minimum, maximum and average length of the time period without usable data is shown in the Table 2 with variable δ_a. The values are in days
2. The probability is shown in the Table 3. The probability can be calculated using parameters of binomial distribution shown in the Table 3
3. Values of *r* in the Table 3 describe the resulting answer
4. Table 3 describes the probability for at least one usable image in the period (P(s=1))

VI. Conclusion

Problem of a real number of usable satellite images, i.e. without cloud cover, is a significant influencer of inland water bodies observation based on remotely sensed data.

The paper proposes a method for evaluation of suitability (usability) of the time series of remotely sensed images for small water bodies observation. It is based on statistical induction. The statistical methods are demonstrated on the case study (observation of small ponds near Pardubice, the Czech Republic) for calculations over the dataset consisting of images coming from Landsat 7 and Landsat 8.

The calculations also show that the real number of usable images is relatively low and it depends on the minimum percentage of clear water surface necessary to next analyses.

The proposed procedure is generalizable. It can be used for further research based on different data sets, i.e. for other satellite data sets with comparable time resolution, for longer time period or different areas of interest (wit focus on small inland water bodies). The calculated values of γ_a show it is a random variable with binomial distribution. It provides possibility to calculate the probabilities and to give quantitative answers for the given questions.

Acknowledgment

This research is supported by University of Pardubice, SGS_2017_17 project.

References

[1] P.L. Brezonik et al, "Factors affecting the measurement of CDOM by remote sensing of optically complex inland waters," Remote Sensing of Environment, vol. 157, pp. 199-215, February 01, 2015.

[2] R. P. Bukata, G. P. Harris, and J. E. Bruton, "The detection of suspended solids and chlorophyll-a utilizing digital multispectral ERTS-1 data," 2nd Can. Symp. Remote Sensing, 1974, pp. 552–564.

[3] R. P. Bukata, G. P. Harris, and J. E. Bruton, "Satellite-observations of water-quality," Transportation engineering journal of asce, pp. 537-554, 1976.

[4] E. M. Middleton, and R. F. Marcell, [Online], "Literature relevant to remote sensing of water quality," 1983, Available at: https://ntrs.nasa.gov/search.jsp?R=19830026142 [cited 2017-02-05].

[5] M. Pásler, J. Komárková, P. Sedlák, "Comparison of possibilities of UAV and Landsat in observation of small inland water bodies," International Conference on Information Society, i-Society 2015, pp. 45-49.

[6] M. Pásler, and J. Komárková, "Utilization of Landsat data for water quality observation in small inland water bodies," International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, vol. 41, pp. 373-377, 2016.

[7] M. Salama, M. Radwan, and R. van der Velde, "A hydro-optical model for deriving water quality variables from satellite images (HydroSat): A case study of the Nile River demonstrating the future Sentinel-2 capabilities," Phys. Chem. Earth, Parts, pp. 224-232, 2012.

[8] P. Sedlák, Z. Szczyrba, E. Kudrnovský, "Spatial temporal changes of land use in postcommunist towns with remote sensing data - case study of Olomouc city", Symposium Remote Sensing of Urban Area. Regensburg, Germany, pp. 176 – 178, 2003.

[9] W. Van Wychen et al., "Variability in ice motion and dynamic discharge from Devon Ice Cap, Nunavut, Canada", Journal of Glaciology, vol. 63, pp. 436-449, June 2017.

[10] S. D. Wang et al, "A Simple Enhanced Water Index (EWI) for Percent Surface Water Estimation Using Landsat Data", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, pp. 90-97, January 2015.

[11] Z. Q. Wang, C.C. Gang, X.L. Li, Y.Z. Chen, and J. L. Li, "Application of a normalized difference impervious index (NDII) to extract urban impervious surface features based on Landsat TM images", International Journal of Remote Sensing, vol. 36, pp. 1055-1069, 2015.

[12] Q. Ying et al., "Global bare ground gain from 2000 to 2012 using Landsat imagery", Remote Sensing of Environment, vol. 194, pp.161-176, June 01, 2017.