Data gap filling using cloud-based distributed Markov Chain Cellular Automata framework for land use and land cover change analysis

Hai Lan [1,2], Kathleen Stewart [1], Zongyao Sha [3], Shujuan Chang [4], and Yichun Xie [5,*]

[1] Department of Geographical Sciences, Univ. of Maryland, College Park, MD, USA
[2] NSF Spatiotemporal Innovation Center, George Mason Univ., Fairfax, VA, USA
[3] School of Remote Sensing & Information Engineering, Wuhan Univ., Wuhan, China
[4] Inner Mongolian Forestry & Rangeland Monitoring and Planning Institute, Hohhot, China
[5] Department of Geography and Geology, Eastern Michigan Univ., Ypsilanti, MI, USA
* Correspondence: yxie@emich.edu; Tel.: +1 734-487-7588

Supplementary Figures and Tables

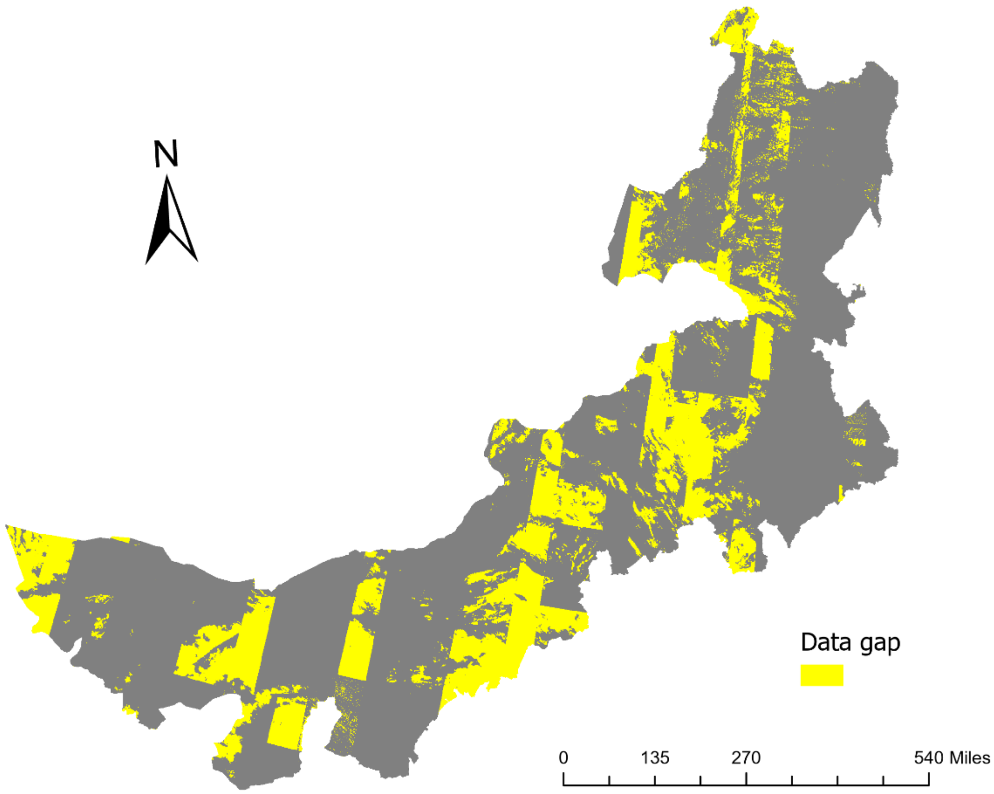

Figure S1. Data gaps in IMAR on August 2016

The portions in yellow are cloud covered areas during August 2016 that correspond to where data gaps exist in IMAR during this period. There are totally ~ 1.28 billion (1281956192) pixels/cells per band in this study area. ~0.49 billion (492899116) pixels are covered by clouds during this month, i.e., the cloud coverage is over 38.45%. In the northern part of the region, where cloud coverage is above 85%, clouds are a major obstacle that potentially impedes LULC research in this region.

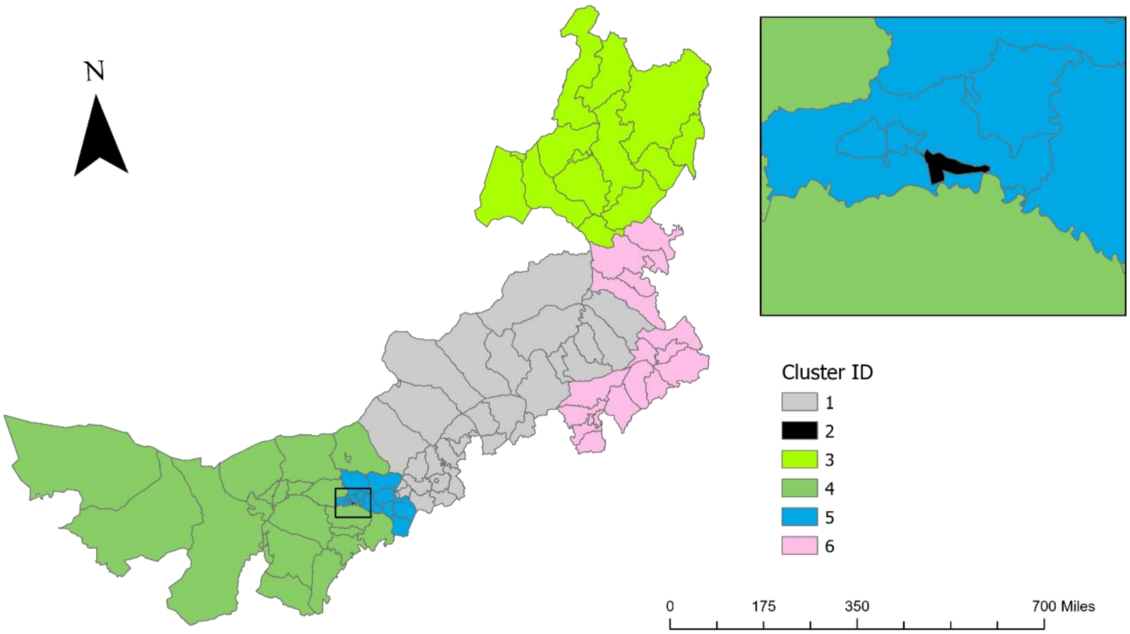

Figure S2. Six clusters generated by county-based multivariate geographic k-medoids clustering analysis

Since Cluster 2 (in black color) is a single banner within Cluster 5 (in blue color), the final number of clusters used in the analysis is 5.

Table S1. Suitability factors

Factors	Definition	Function Shape	Control Point (s) / SB	
DEM	Elevation of whole IMAR (m)	MDJ	SSS 1	690, 1143
			SSS 2	387, 973
			SSS 3	1024, 1672
			SSS 4	1414, 1969
			SSS 5	1316, 1852
SLOPE	Slope calculated by DEM (degree)	MDJ	SSS 1	6.1, 20.5
			SSS 2	3.9, 17.3
			SSS 3	4.1, 14.9
			SSS 4	11.1, 27.7
			SSS 5	4.5, 15.8
DisRoad	Distance to roads	Linear	N/A	
DisRail	Distance to railways	Linear	N/A	
PPT	Yearly average precipitation (mm)	SS	SSS 1	389.35
			SSS 2	419.88
			SSS 3	276.29
			SSS 4	361.9
			SSS 5	152.4
TEMP	Yearly average temperature (centigrade)	SS	SSS 1	-1.58
			SSS 2	5.97
			SSS 3	3.28
			SSS 4	6.5
			SSS 5	8.38
POP	Population density per each county (people per km^2)	MDJ	SSS 1	5.1, 37
			SSS 2	37.8, 338.1
			SSS 3	8.1, 194.5
			SSS 4	52.0, 379.4
			SSS 5	1.8, 57.5
LS	Livestock density per each county (livestock per km^2)	MDJ	SSS 1	11.1, 74.9
			SSS 2	115.9, 474.6
			SSS 3	36.9, 240.6
			SSS 4	116.4, 521.2
			SSS 5	14.2, 138.9
CP	Compensation policy	Linear	0,1,2	
WA	Water area	Boolean	N/A	
LU	Human land use	Boolean	N/A	

Restraint factors affected LULC change with specific functions and control points. For example, DisRoad and DisRail represented the distance to roads and railways. The closer roads and railways were (short distances), the less chance there was that the original LULC classes would be converted into grassland classes. In this study, the effect of the distance to road and railways was defined as a linear function. Relating to animal grazing practices in the study area, the compensation policy included three practices: normal grazing (not participated in), balanced grazing, and forbidden grazing. This was a county-based policy that allowed herdsman to get compensation for balanced grazing or to halt grazing for specific time periods. Some counties in IMAR did not participate in this program. In this study, a linear function was defined for this factor by considering that with less grazing, there would be a higher chance that grasslands may recover from other types of LULC classes. Water areas (WA) and human land use areas (LU) were two constraint factors that were not expected to be changed into grassland classes.

Different function shapes were also applied to optimize the suitability value calculations. In this study, four types of functions were applied in MCE: Monotonically decreasing J-shaped function (MDJ), linear, symmetric sigmoidal (SS), and boolean. For example, in Sub-Simulation Space (SSS) 1, a monotonically decreasing J-shaped function was used to calculate suitability values for the DEM layer. Control points defined the position and accurate shape of the function curve in the Cartesian Coordinate System. Shape was used for cases such that, for example, the potential suitability of grassland was similar for areas with less than 690m elevation. Suitability gradually decreased however, as elevation increased from 690m to 1143m. Land was deemed unsuitable for grassland at elevations higher than 1143m. Another example was a symmetric sigmoidal function that was used to calculate the values for the yearly average temperature layer. To simplify calculations, we set -1.58 ºC as the uppermost value for average yearly temperature according to the average yearly temperature from 2010 to 2016, which represented that the grassland suitability value was highest at -1.58 ºC, and gradually decreased with lower or higher temperatures. After processing all eleven suitability factors, a suitability grid set was generated with assigned weights for each suitability grid. In this study, we applied equal weights to all eleven factors to simplify the modeling and computing.

Table S2. Transition Matrix

	SSS	TMS	TS	TDS	TSD	TD	LM	MM	MA	WA	NG
TMS	1	71.87	8.03	N/A	N/A	N/A	2.12	0.34	0.09	0	17.54
	2	66.09	0	N/A	N/A	N/A	2.33	0.87	0.02	0.07	30.62
	3	58.59	22.35	0	0	N/A	3.01	0.53	0	0.17	15.36
	4	9.66	48.6	0	N/A	N/A	0	0	0	0.5	41.24
	5	16.91	76.58	0	0.35	0	0	0	0	0	6.16
TS	1	9.97	79.16	N/A	N/A	N/A	10.7	0	0.09	0.08	0
	2	0	65.43	N/A	N/A	N/A	3.75	0	0	0.26	30.57
	3	10.34	34.57	22.63	0	N/A	14.66	0.01	0	0.4	17.39
	4	0.24	60.14	0.03	N/A	N/A	0	0	0	2.11	37.48
	5	0	73.37	11.71	0	0	0.27	0	0	0.98	13.66
TDS	1	N/A									
	2	N/A									
	3	0	22.76	67.73	7.03	N/A	2.48	0	0	0	0
	4	0	10.65	86.34	N/A	N/A	0	0	0	0.32	2.69
	5	0	28.95	42.62	15.28	1.92	3.71	0.01	0	0.24	7.26
TSD	1	N/A									
	2	N/A									
	3	0.15	0	33.17	63.43	N/A	3	0	0	0	0.23
	4	N/A									
	5	0	0	21.89	52.12	25.63	0	0.04	0	0.32	0
TD	1	N/A									
	2	N/A									
	3	N/A									
	4	N/A									
	5	0	0	0	22.05	31.87	11.45	0	0.58	0	34.05
LM	1	1.01	4.51	N/A	N/A	N/A	54.59	1.75	1.18	2.31	34.65
	2	1.87	9.23	N/A	N/A	N/A	44.93	0	0.25	5.28	38.44
	3	6.62	29.2	2.13	2.46	N/A	50.11	0	0.1	3.24	6.14
	4	0	6.79	0	N/A	N/A	17.99	0	0	11.03	64.19
	5	0	7.42	16.44	4.41	11.44	30.38	0	0.21	5.99	23.71
MM	1	22.28	0	N/A	N/A	N/A	3.97	55.88	0	0	17.87
	2	41.08	0	N/A	N/A	N/A	2.83	31.27	0	0	24.82
	3	64.1	0	0.6	0.06	N/A	0.23	17.77	0.01	0	17.22
	4	27.8	0.82	0	N/A	N/A	0.03	69.52	0	0	1.84
	5	0	0	0	0	0	0	74.55	0	0	25.45
MA	1	2.48	3.2	N/A	N/A	N/A	47.45	0	43.31	3.56	0
	2	6.31	0.35	N/A	N/A	N/A	37.12	0	20.75	0	35.47
	3	2.83	0	0	0	N/A	96.91	0	0	0.26	0
	4	0	0	0	N/A	N/A	56.7	0	0	41.88	1.42
	5	0	0	0	0.8	0	0	0	35.23	59.18	4.79
WA	1	0	2.22	N/A	N/A	N/A	19.72	0	0.65	77.41	0
	2	0	0	N/A	N/A	N/A	22.13	0	1.15	44.55	32.17
	3	0	0.93	0	0	N/A	33.11	0	0.12	65.83	0
	4	0	0	0	N/A	N/A	37.8	0	0	47.51	14.68
	5	0	1.51	4.13	0	45.03	24.73	0	0.87	23.74	0
NG	1	9.9	0	N/A	N/A	N/A	32.12	4.54	0	0.71	52.73
	2	18.12	16.43	N/A	N/A	N/A	3.96	0.19	0	1.5	59.81
	3	18.07	29.78	0	0	N/A	0.59	0.66	0.02	1.33	49.55
	4	0.19	14.34	0.01	N/A	N/A	1.38	0.01	0.13	1.79	82.14
	5	0.11	13	12.32	0.34	31.96	6.56	0.05	0	2.41	33.25

The transition matrix for the ten LULC classes for the IMAR region was calculated (ST. 2). In this table, the diagonal values for each sub-simulation space (SSS) represented the probability that certain LULC classes that did not change during the training period. Other values in this table show the probability that certain LULC classes were changed to other types of LULC classes. The N/A value in this table represented the case where there was no such class in this SSS. During the two paired training periods 2000 to 2010, and 2010 to 2016, the most stable LULC classes were water area (WA) and non-grassland area (NG). They were considered as Boolean types of constrain factors, which were expected to be fixed and not change during the simulation. Other types of LULC classes can be changed into those two classes given a certain probability. However, in reality, it was quite rare that grassland classes changed into water areas. On other hand, rapid increases in human activities resulted in grassland classes being changed into NG classes, especially to farmland and human land use (LU). For example, cities and roads were expanded rapidly in IMAR. Over grazing also happened in some areas during past few decades. All those cases may lead to grassland degradation or even the disappearance of grassland especially in some sensitive areas. By assessing this transition matrix, we found that Temperate Meadow-steppe (TMS), temperate steppe (TS), and Lowland Meadow (LM) were changed into NG with a relatively high probability (typically over 30%) especially in SSS 2 and SSS 4, which were high population density and high livestock density regions respectively. We also found that some transitions happened between neighboring types of grassland classes. For example, TMS class showed a probability of approximately 76.58% to change into TS in SSS 5, which is a very dry area. In other words, TMS class areas that require relatively high moisture environments may be less likely to be stay as dry areas over time.

The transition matrix represented the analysis results from a mathematical perspective. During the simulation process, the results that were closer to an actual change mechanism were generated by integrating multi-types of impact factors, weighting each of them with different functions and assigning different neighborhood weights.

Table S3. Overall accuracy assessment without using sub-region strategy with 2 years data

	TMS	TS	TDS	TSD	TD	LM	MM	MA	WA	NG
Producer's Accuracy	79.30	77.65	73.14	74.11	82.35	81.87	76.22	29.79	40.91	87.37
User's Accuracy	85.59	81.67	90.30	81.92	92.17	74.40	77.86	56.00	26.73	77.44
Overall Accuracy	81.23									
Kappa	0.76									

Table S4. Overall accuracy assessment using sub-region strategy with 2 years data

	TMS	TS	TDS	TSD	TD	LM	MM	MA	WA	NG
Producer's Accuracy	84.02	87.10	80.59	80.81	83.37	83.38	82.35	38.10	42.75	88.30
User's Accuracy	88.61	85.01	92.59	83.48	92.14	74.24	80.26	39.02	26.79	85.00
Overall Accuracy	85.23									
Kappa	0.81									

Table S5. Overall accuracy assessment without using sub-region strategy with 3 years data

	TMS	TS	TDS	TSD	TD	LM	MM	MA	88.165	NG
Producer's Accuracy	81.10	82.19	76.76	73.85	83.78	85.65	81.12	38.10	46.28	89.11
User's Accuracy	88.25	82.21	92.25	82.20	92.28	75.25	78.91	59.26	27.45	83.11
Overall Accuracy	83.97									
Kappa	0.79									

Table S6. Overall accuracy assessment by using sub-region strategy with 3 years data

	TMS	TS	TDS	TSD	TD	LM	MM	MA	WA	NG
Producer's Accuracy	87.20	91.88	89.71	89.23	89.51	84.14	86.67	94.12	59.76	85.93
User's Accuracy	89.04	89.13	93.20	84.89	98.05	77.79	85.53	28.07	23.67	86.89
Overall Accuracy	88.16									
Kappa	0.85									

Table S7. Sub-regions accuracy assessment

		TMS	TS	TDS	TSD	TD	LM	MM	MA	WA	NG
SSS 1	Producer's Accuracy	85.61	98.81	N/A	N/A	N/A	85.71	82.76	100	55.00	96.47
	User's Accuracy	93.70	99.10	N/A	N/A	N/A	87.64	96.00	100	100	93.57
	Overall Accuracy	93.82									
	Kappa	0.89									
SSS 2	Producer's Accuracy	81.30	74.88	N/A	N/A	N/A	70.87	100	100	31.25	88.41
	User's Accuracy	81.63	80.26	N/A	N/A	N/A	76.84	80.0	90.0	41.67	85.74
	Overall Accuracy	83.50									
	Kappa	0.71									
SSS 3	Producer's Accuracy	93.71	93.55	98.66	95.74	N/A	78.88	100	100	54.55	68.59
	User's Accuracy	90.24	94.44	99.66	100	N/A	92.56	100	90.0	54.55	65.52
	Overall Accuracy	89.78									
	Kappa	0.86									
SSS 4	Producer's Accuracy	90.91	71.54	100	N/A	N/A	30.77	90.91	100	15.0	87.34
	User's Accuracy	100	69.96	90.0	N/A	N/A	80.0	100	100	30.0	84.78
	Overall Accuracy	80.84									
	Kappa	0.60									
SSS 5	Producer's Accuracy	100	72.12	95.42	95.43	98.39	79.17	100	90.91	40.0	78.90
	User's Accuracy	20.0	84.40	88.56	99.40	95.10	75.0	70.0	100	54.55	83.88
	Overall Accuracy	89.74									
	Kappa	0.86									