Frontiers in Landsat Change Detection: Assessing the Quality of Time-Series Maps

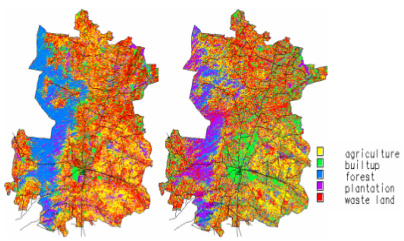
Warren B. Cohen, US Forest Service

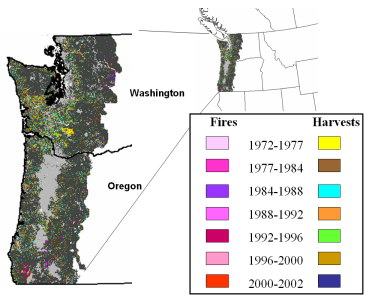
Zhiqiang Yang, Robert E. Kennedy, Peder Nelson

Oregon State University

Landsat Data Product Workshop, Boston University – 27-29 October 09

Change Detection: “Where Do We Come From, Where Do We Go*





Healey et al. 2009

1972

1980

1990

2000

2008

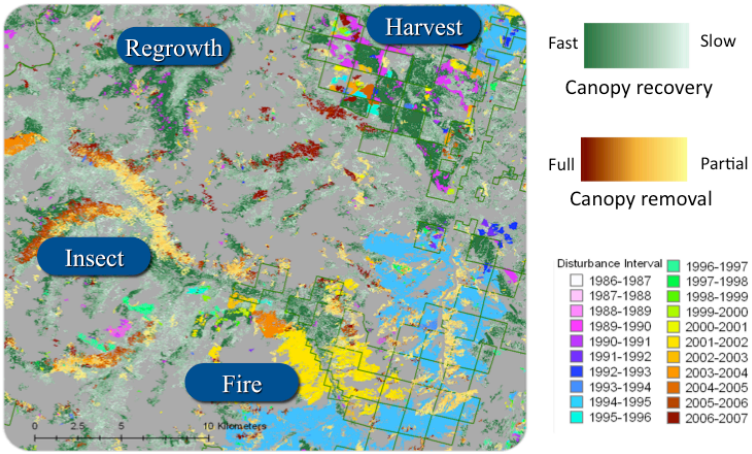
1 Landsat scene, 2 image dates

Slow, incremental progress

100s Landsat scenes, 2 image dates (Masek et al, LEDAPS) ***or*** handfuls of scenes and dates (eg, Healey et al)

A new frontier

2009 and beyond – 100s Landsat scenes ***and*** up to nearly 40 years of imagery, and lots more information



From RE Kennedy

* From the great American Old-Time tune Cotton-Eyed Joe

New Frontiers for Monitoring Change with Landsat Data

- Why a new frontier?
  - Opening of the USGS Landsat archive with a single, high-quality processing recipe
  - Tremendous need for more detailed information
- What is enabling the passage?
  - Sophisticated, automated algorithms (eg, LandTrendr – Kennedy et al; VCT – Huang et al)
- What else is needed?
  - Calibration & validation methods/protocols that fill the <u>temporal and spatial voids</u> associated with extant ground or airphoto data

Free Access to Landsat Imagery

WE ARE ENTERING A NEW ERA IN THE LANDSAT Program, the oldest and most venerable of our Earth-observing satellite programs. With little fanfare, the U.S. Geological Survey (USGS) has begun providing imagery for free over the Internet. Throughout the history of the Landsat Program, the cost and access to imagery has always limited our ability to study our planet and the way it is changing. Beginning with a pilot program to provide "Web-enabled" access to Landsat 7 images of the United States that were collected between 2003 and this year, the USGS now plans to provide top-quality image products for free upon request for the entire U.S. archive, including over 2 million images back to Landsat 1 (1972) [for details and schedules, see (*1*)]. The release by NASA and the USGS in January 2008 of a new Landsat Data Distribution Policy (*2*) was a key step to this goal. Free imagery will enable reconstruction of the history of Earth's surface back to 1972, chronicling both anthropogenic and natural changes during a time when our population doubled and the impacts of climate change became noticeable.

Free image. This Landsat 5 image of the southeastern corner of the Black Sea is part of the general U.S. archive that will be accessible for free under the new USGS policy.

THE LANDSAT SCIENCE TEAM: CURTIS E. WOODCOCK,[1*] RICHARD ALLEN,[2] MARTHA ANDERSON,[3] ALAN BELWARD,[4] ROBERT BINDSCHADLER,[5] WARREN COHEN,[6] FENG GAO,[3] SAMUEL N. GOWARD,[7] DENNIS HELDER,[8] EILEEN HELMER,[9] RAMA NEMANI,[10] LAZAROS OREOPOULOS,[5] JOHN SCHOTT,[11] PRASAD S. THENKABAIL,[12] ERIC F. VERMOTE,[13] JAMES VOGELMANN,[14] MICHAEL A. WULDER,[15] RANDOLPH WYNNE[16]

www.sciencemag.org SCIENCE VOL 320 23 MAY 2008
Published by AAAS

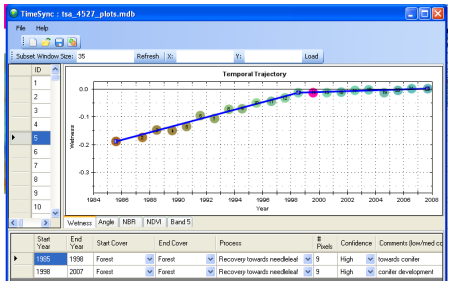
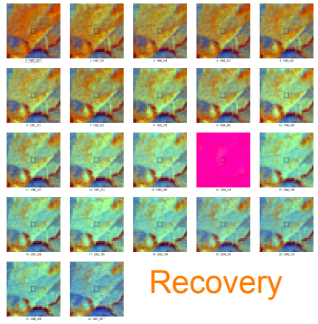
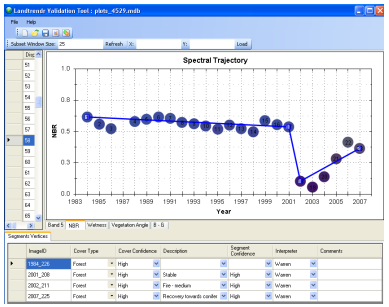


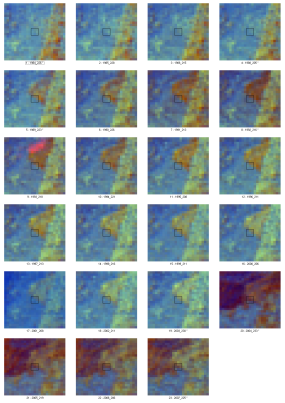
Image "chip" series



Recovery

Spectral trajectory



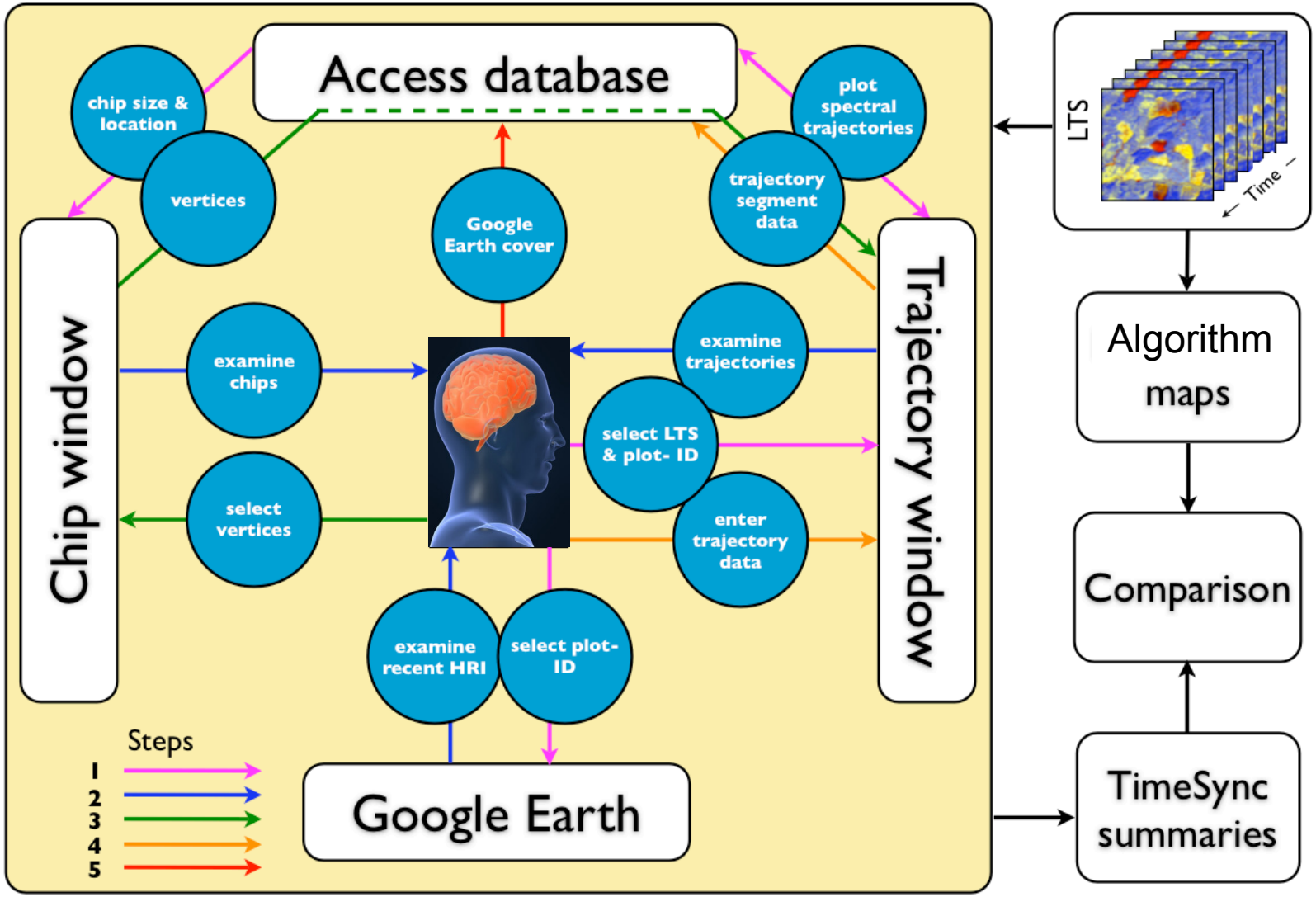


Disturbance

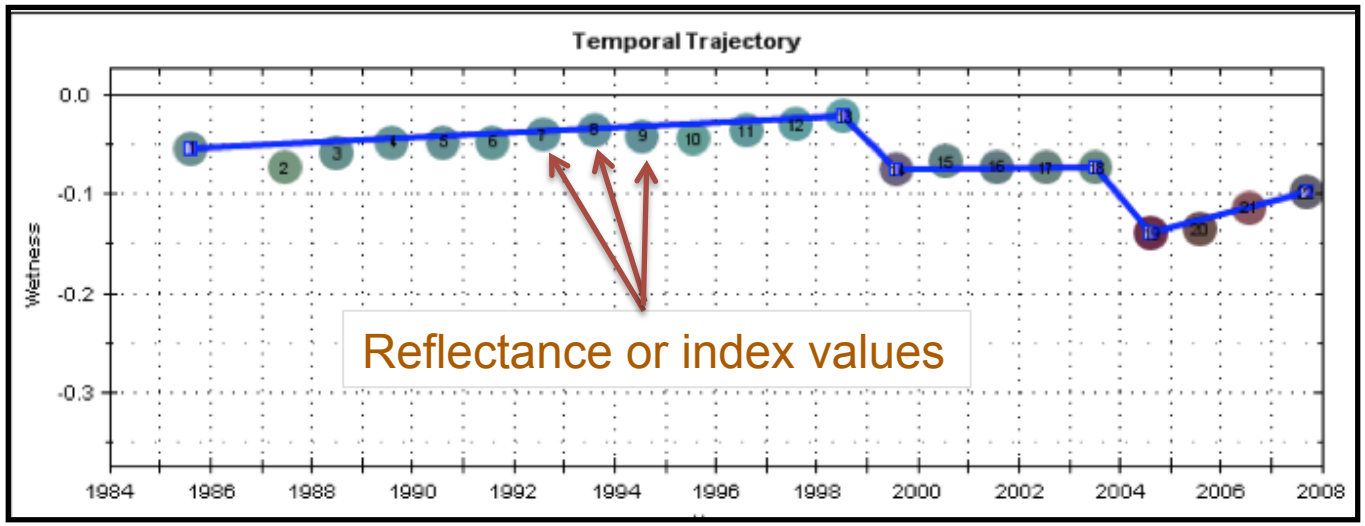
Temporal and Spatial Voids

- Change can occur anywhere at anytime!
- Where are the reference data?
  - Potential requirement for any given location at any time during the period of evaluation (1972-future)
  - At best, they may exist in some locations for some dates, but mostly missing
  - Sampling can be biased
  - Extant data can be costly to obtain, assemble, and evaluate
- Can we use the time series itself, in photo-interpretation mode?
  - Above problems would largely dissipate
  - Precedent exists for interval-based, high intensity forest disturbances: Cohen et al. 1998, Hayes and Sader 2001, Masek et al. 2008, and others)
  - Extant data can still be used, as needed
- TimeSync
  - Developed as one solution to address these problems

TimeSync: Synching human and automated interpretations of Landsat time series -
Enhanced "photo-interpretation" of image time series at specific locations ("plots")

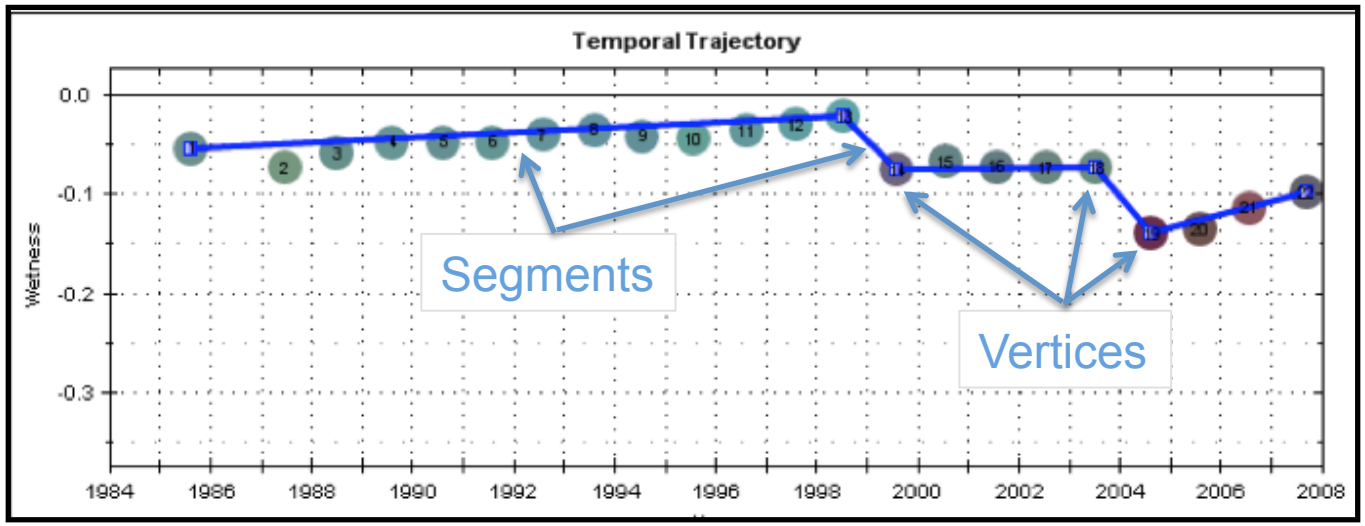


Annual Landsat Trajectory: Change Elements



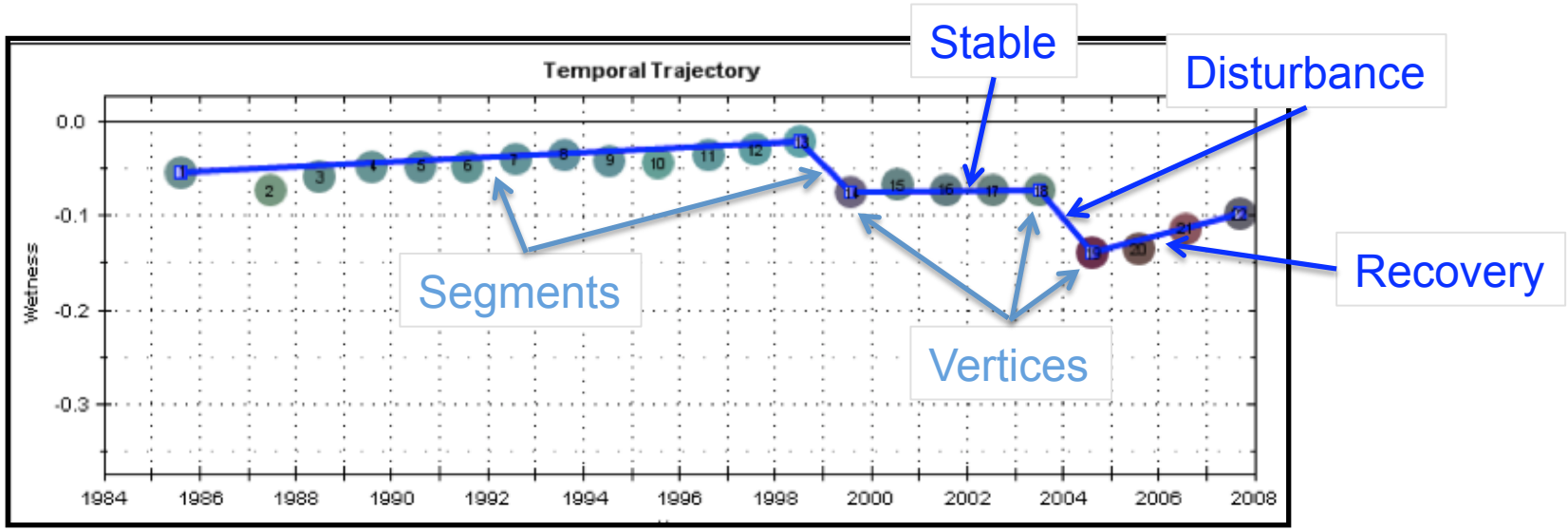
- Annual spectral response in favorite band or index

Annual Landsat Trajectory: Change Elements



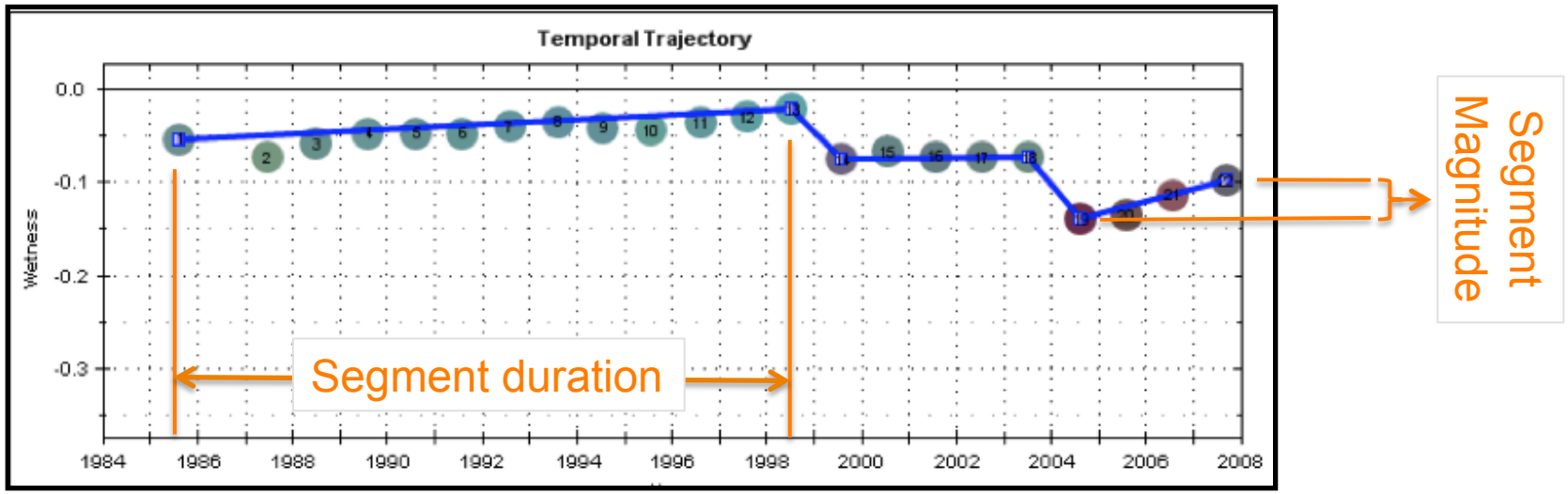
- Annual spectral response in favorite band or index
- ~ Linear segments with start and end vertex dates

Annual Landsat Trajectory: Change Elements

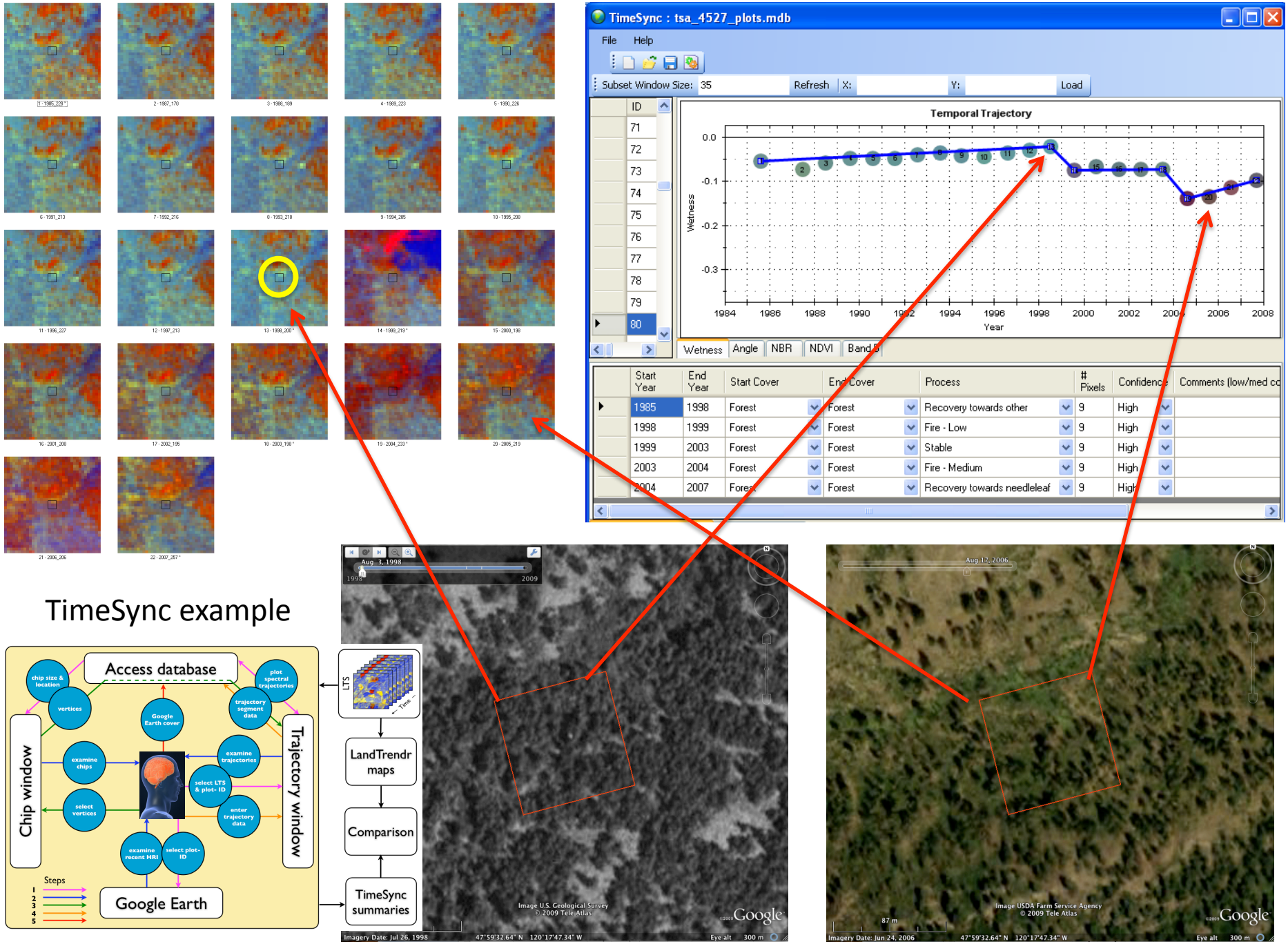


- Annual spectral response in favorite band or index
- ~ Linear segments with start and end vertex dates, and labels (stable, disturbance, recovery)

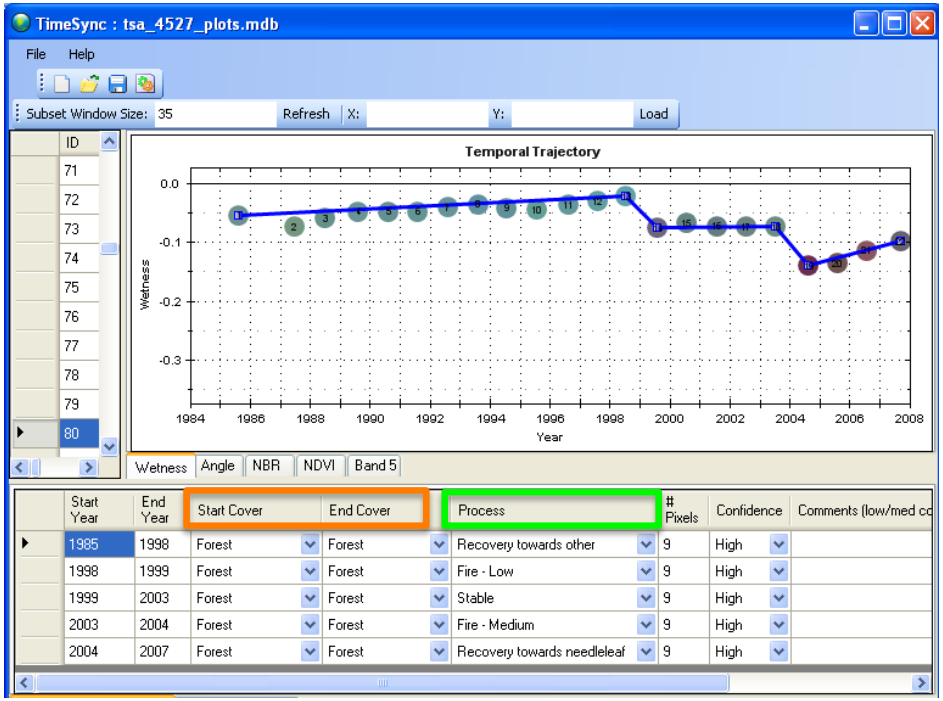
Annual Landsat Trajectory: Change Elements



- Annual spectral response in favorite band or index
- ~ Linear segments with start and end vertex dates, and labels (stable, disturbance, recovery)
- Segment duration (years) & spectral change magnitude in band or index


TimeSync : tsa_4527_plots.mdb
File
Help
Subset Window Size: 35
Refresh
X:
Y:
Load
ID
71
72
73
74
75
76
77
78
79
80
Temporal Trajectory
Wetness
Year
Wetness
Angle
NBR
NDVI
Band 5
Start Year
End Year
Start Cover
End Cover
Process
Pixels
Confidence
Comments (low/med co
1985 1998 Forest Forest Recovery towards other 9 High
1998 1999 Forest Forest Fire - Low 9 High
1999 2003 Forest Forest Stable 9 High
2003 2004 Forest Forest Fire - Medium 9 High
2004 2007 Forest Forest Recovery towards needleleaf 9 High
TimeSync example
Access database
Chip window
Trajectory window
Google Earth
Steps
LTS
LandTrendr maps
Comparison
TimeSync summaries
Image U.S. Geological Survey
© 2009 Tele Atlas
Image USDA Farm Service Agency
© 2009 Tele Atlas
Google

Selectable Cover/Use and Process



Vertices

Land Use/ Cover Class

Forest
Needleleaf forest
Broadleaf forest
Mixed forest
Woodland
Shrubland
Grassland
Agriculture
Wetland
Urban
Snow/ice
Barren
Water
Other

Segments

Process*

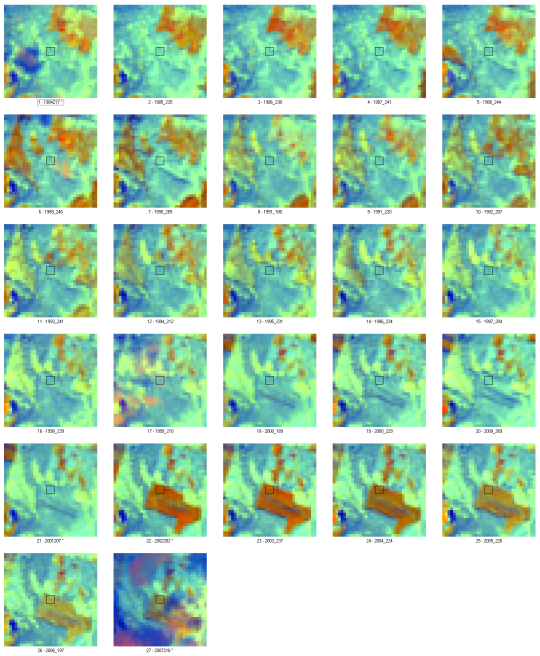
Stable
Recovery
Recovery towards needleleaf
Recovery towards broadleaf
Woody encroachment
Harvest
Fire
Pathogen
Blowdown
Other

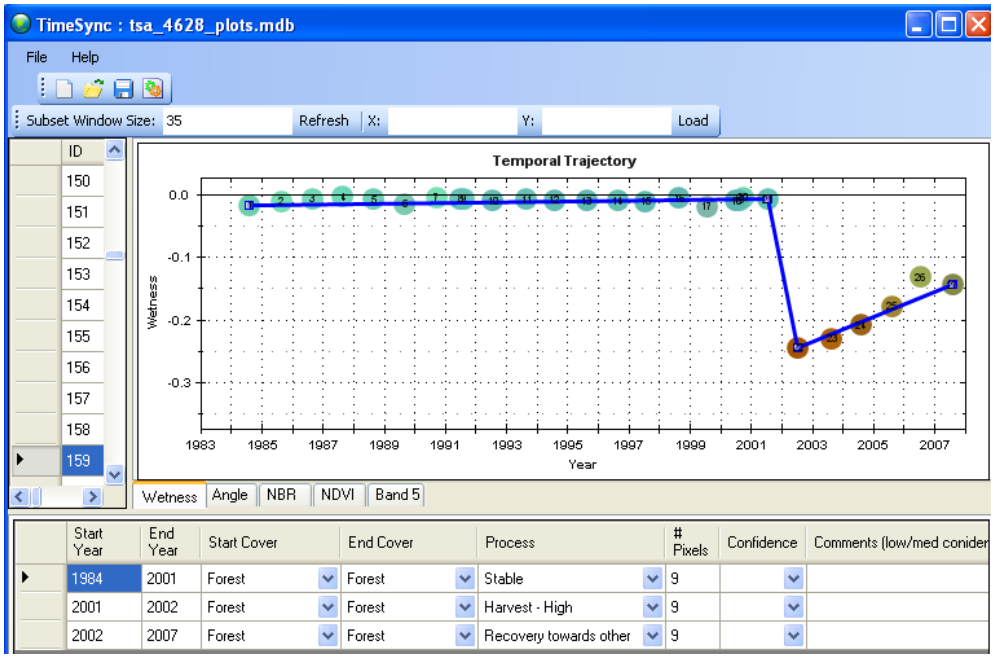
* Intensity (L,M,H) for disturbances

- Any band or index
- # pixels affected
- Confidence and comment

TimeSync Validation

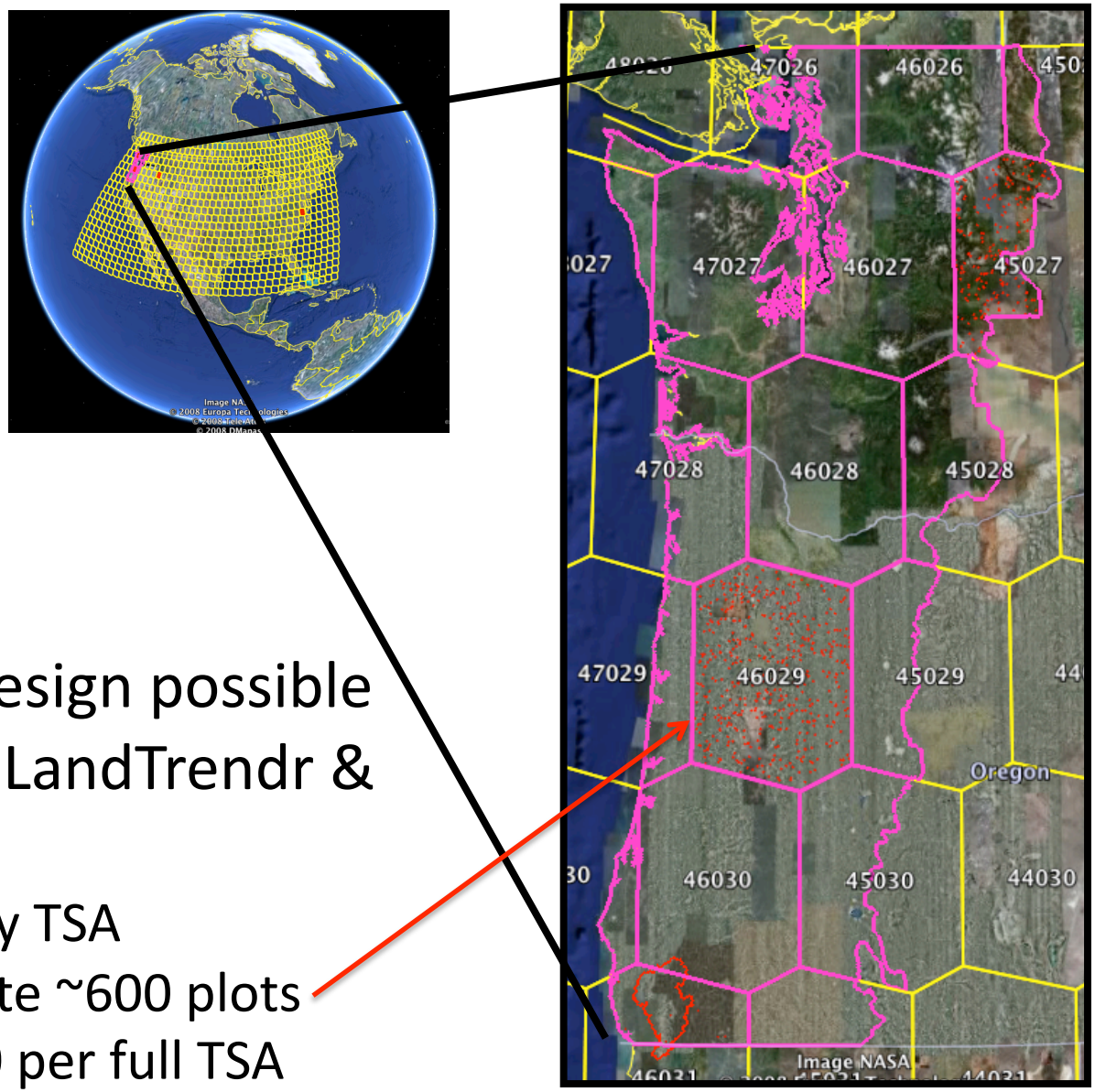
- Change described in terms of segment categories and duration, start vertex date, and agent & relative intensity (for disturbance)





- How to compare TimeSync observations and algorithm/map output?

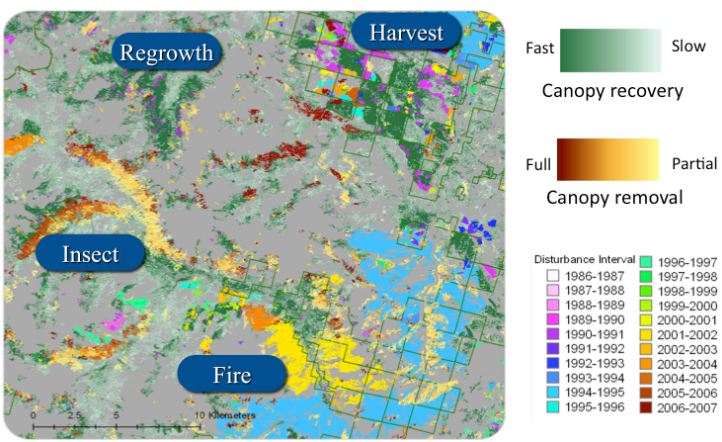
TimeSync: Validation Sample Design.....



- Any statistical design possible
- Design used for LandTrendr & the NWFP:
  - Stratification by TSA
  - Randomly locate ~600 plots
  - Interpret ~ 200 per full TSA

Example of TimeSync v. Automated Algorithm over 4 TSAs

- Raw materials
  - TimeSync segment observations of detailed type, start date and duration, & intensity (disturbance)
  - Time series map segments of type, start and duration, & magnitude



From Kennedy

Segment category by intensity (disturbance only)	Number	Per disturbance type	Per disturbance intensity	Percent
Harvest - High	50	-	-	-
Harvest - Medium	38	-	-	-
Harvest - Low	38	-	-	-
Harvest	-	126	-	0.72
Fire - High	7	-	-	-
Fire - Medium	19	-	-	-
Fire - Low	5	-	-	-
Fire		31	-	0.18
Pathogen - High	0	-	-	-
Pathogen - Medium	6	-	-	-
Pathogen - Low	8	-	-	-
Pathogen		14	-	0.08
Other - High	2	-	-	-
Other - Medium	0	-	-	-
Other - Low	3	-	-	-
Other	-	5	-	0.03
Disturbance type total	-	176	-	1.00
High	-	-	59	0.34
Medium	-	-	63	0.36
Low	-	-	54	0.31
Disturbance intensity total	-	-	176	1.00
Disturbance total	176	-	-	0.25
Recovery total	250	-	-	0.36
Stable total	277	-	-	0.39
Segment total	703	-	-	1.00

Segmentation summary – 388 forested [plots

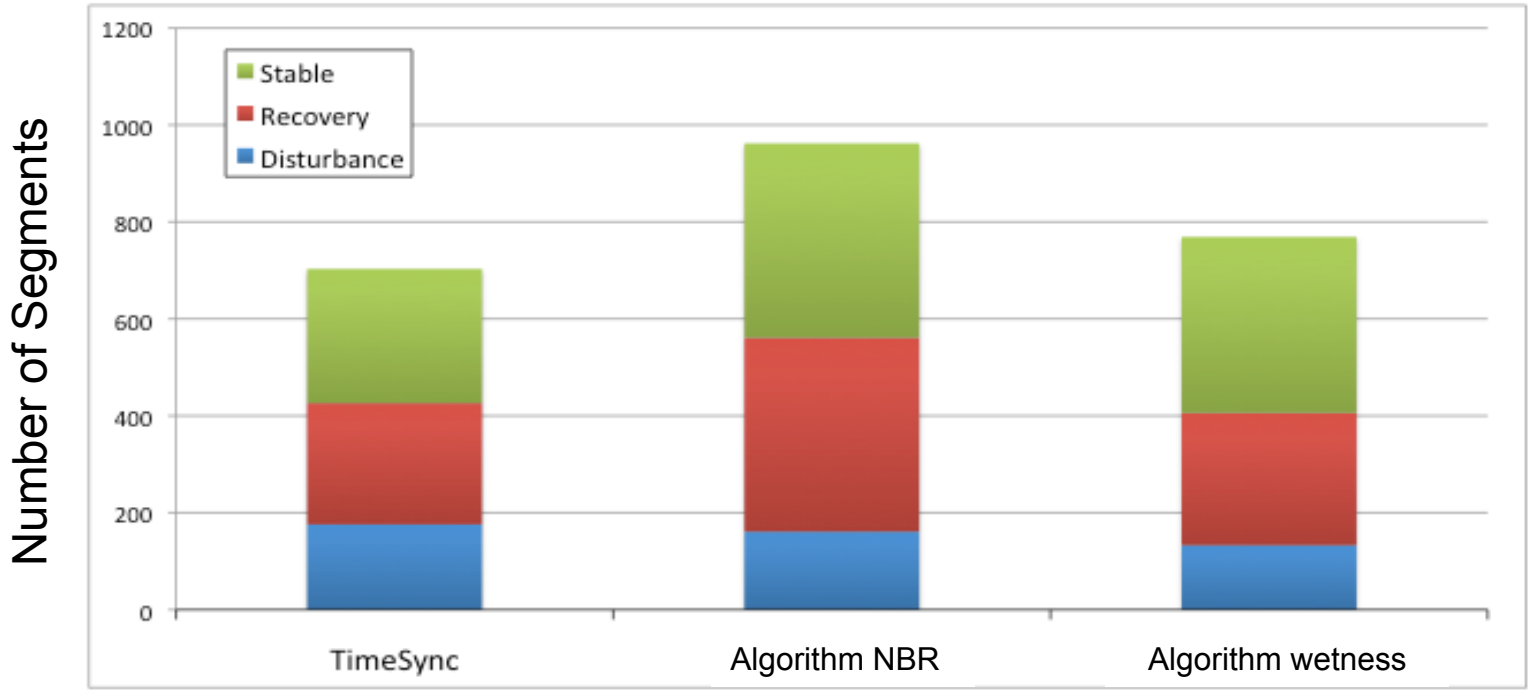
TimeSync v. Algorithm Comparisons – Many Possibilities

- Segment summary distributions
  - Number of segments per plot & across plots
  - Number by category (disturbance, recovery, stable)
- Overall match scores
  - Category label match by year across trajectory, overall and by category
    - All three categories; Disturbance v. No Disturbance
  - Examine worst cases using TimeSync and algorithm trajectory plot
- Vertex match score matrices
  - Segment vertex match by category
    - Restricted and relaxed (fuzzy) precision
  - Examine all false negatives and positives (as above), and spectral change magnitude histograms for agreement and disagreement
- Proportion in each major category across plots and time

TimeSync v. Algorithm Comparisons – Many Possibilities

- Segment summary distributions
  - Number of segments per plot & across plots
  - Number by category (disturbance, recovery, stable)
- Overall match scores
  - Category label match by year across trajectory, overall and by category
    - All three categories; Disturbance v. No Disturbance
  - Examine worst cases using TimeSync and algorithm trajectory plot
- Vertex match score matrices
  - Segment vertex match by category
    - Restricted and relaxed (fuzzy) precision
  - Examine all false negatives and positives (as above), and spectral change magnitude histograms for agreement and disagreement
- Proportion in each major category across plots and time

Segment Category Summaries



- TimeSync fit ~ 700 segments over 388 forested plots; algorithm run on two different indices
- Algorithm using NBR slightly underfit disturbance but overfit recovery and stable
- Algorithm using wetness underfit disturbance, but overfit other to lesser extent
- NRB more sensitive index for disturbance, but comes at expense of overfitting non-disturbance categories

TimeSync v. Algorithm Comparisons – Many Possibilities

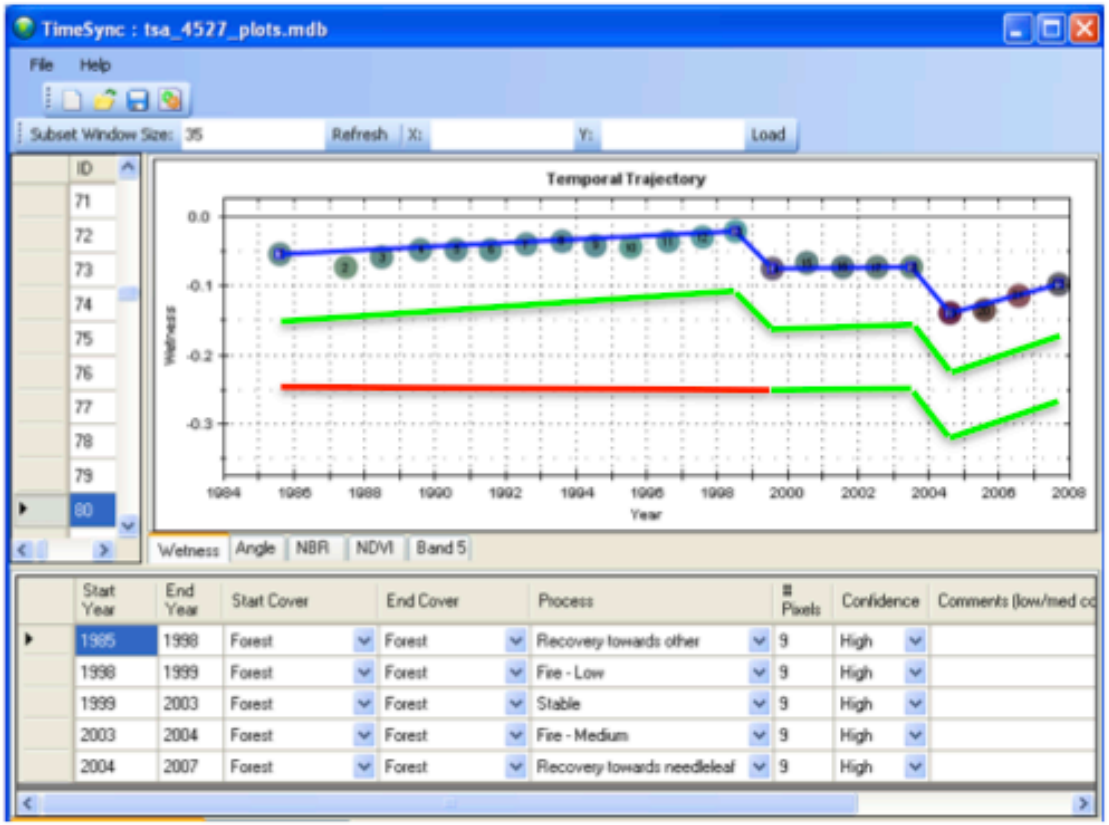
- Segment summary distributions
  - Number of segments per plot & across plots
  - Number by category (disturbance, recovery, stable)
- Overall match scores
  - Category label match by year across trajectory, overall and by category
    - All three categories; Disturbance v. No Disturbance
  - Examine worst cases using TimeSync and algorithm trajectory plot
- Vertex match score matrices
  - Segment vertex match by category
    - Restricted and relaxed (fuzzy) precision
  - Examine all false negatives and positives (as above), and spectral change magnitude histograms for agreement and disagreement
- Proportion in each major category across plots and time

Overall Match Scoring—How It Works

Match
Mismatch
Range: 0 – 100%

Plot-level Score

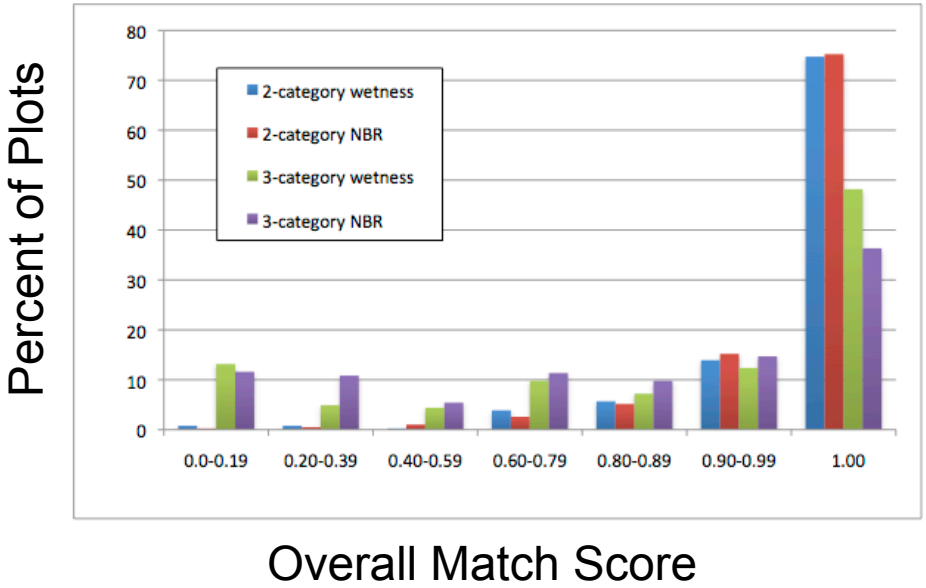
Map1 = 100%
Map2 = 41%



TS
Map1
Map2

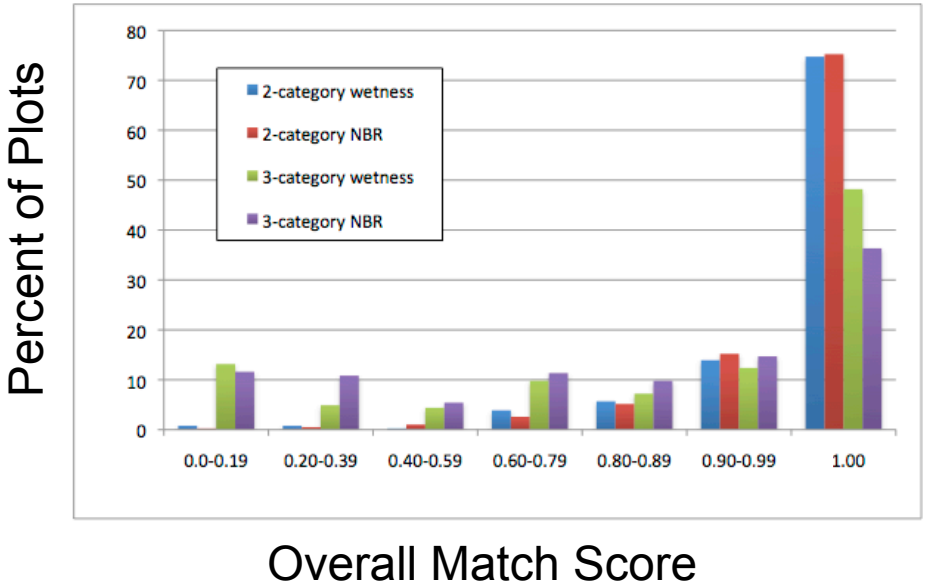
- Segment category labels assessed for agreement at every year of the trajectory
- Plot-level score is proportion of years of agreement over total number of years

Overall Match Scoring—Distributions Across Plots



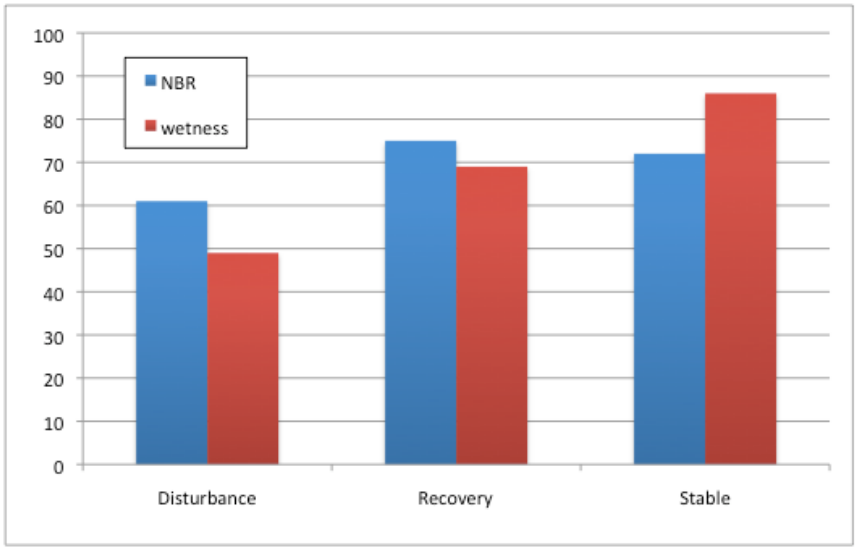
- For disturbed v. not disturbed, ~75% plots completely agree with TimeSync for both indices
- Lesser agreement for all three categories
  - e.g., mean score equals 73.3% and 96.5%, respectively, for 3- and 2-categories for NBR

Overall Match Scoring—Distributions Across Plots



- For disturbed v. not disturbed, ~75% plots completely agree with TimeSync for both indices
- Lesser agreement for all three categories
  - Mean score equals 73.3% and 96.5%, respectively, for 3- and 2-categories for NBR

Mean OMS by category



- Better agreement for non-disturbance categories
- NBR has better agreement for disturbance, but less agreement for stable

TimeSync v. Algorithm Comparisons – Many Possibilities

- Segment summary distributions
  - Number of segments per plot & across plots
  - Number by category (disturbance, recovery, stable)
- Overall match scores
  - Category label match by year across trajectory, overall and by category
    - All three categories; Disturbance v. No Disturbance
  - Examine worst cases using TimeSync and algorithm trajectory plot
- Vertex match score matrices
  - Segment vertex match by category
    - Restricted and relaxed (fuzzy) precision
  - Examine all false negatives and positives (as above), and spectral change magnitude histograms for agreement and disagreement
- Proportion in each major category across plots and time

Vertex Match Score Matrices (restricted)

- Searching the plot trajectories for segment start dates and tallying results into table yields agreement/error matrix

	Algorithm NBR				
TimeSync	*Disturbance*	*Recovery*	*Stable*	*None*	*LT Omission*
Disturbance	103	2	6	64	**0.41**
Recovery	0	160	19	70	**0.36**
Stable	6	39	188	44	**0.32**
None	46	191	188	7911	**0.05**
LT Commission	**0.34**	**0.59**	**0.53**	**0.02**	

- Here, when TimeSync first observed a disturbance (start vertex), the algorithm missed that start date 41% of the time
- This is similar to standard assessment of a single change map based on intervals, but here
  - We added other segment types (recovery and stable), and
  - Based the assessment on a collection of 1-year interval maps

Vertex Match Score Matrices (relaxed)

Algorithm NBR

TimeSync	*Disturbance*	*Recovery*	*Stable*	*None*	*LT Omission*
Disturbance	126	3	7	40	**0.28**
Recovery	0	214	20	16	**0.14**
Stable	6	42	204	25	**0.26**
None	29	141	170	7994	**0.04**
LT Commission	**0.22**	**0.47**	**0.49**	**0.01**	

Overall agreement = 94.5%
Kappa = 68.1%

Algorithm wetness

TimeSync	*Disturbance*	*Recovery*	*Stable*	*None*	*LT Omission*
Disturbance	116	4	7	49	**0.34**
Recovery	2	184	40	24	**0.26**
Stable	9	22	223	23	**0.19**
None	6	63	93	8172	**0.02**
LT Commission	**0.13**	**0.33**	**0.39**	**0.01**	

Overall agreement = 96.2%
Kappa = 75.4%

- Relaxing the match precision by 1 year, significantly improves agreement
  - From 41% omission and 34% commission on disturbance (NBR)
- As with other assessment strategies, wetness does a poorer job with disturbance, but does a better job overall
- Much of the disagreement is distinction between recovery and stable

Vertex Match Score Matrices (relaxed)

Algorithm NBR

TimeSync	*Disturbance*	*Recovery / Stable*	*None*	*LT Omission*
Disturbance	126	0	50	**0.28**
Recovery / Stable	5	477	28	**0.06**
No match	30	44	8277	**0.01**
LT Commission	**0.22**	**0.08**	**0.01**	

Overall agreement = 98.3%
Kappa = 87.8%

Algorithm wetness

TimeSync	*Disturbance*	*Recovery / Stable*	*None*	*LT Omission*
Disturbance	116	1	59	**0.34**
Recovery / Stable	9	466	35	**0.09**
No match	8	43	8300	**0.01**
LT Commission	**0.13**	**0.09**	**0.01**	

Overall agreement = 98.3%
Kappa = 87.6%

- Relaxing by 1 year and aggregating recovery and stable into single class, further improves agreement
  - From 68.1% (NBR) and 75.4% (wetness) Kappa

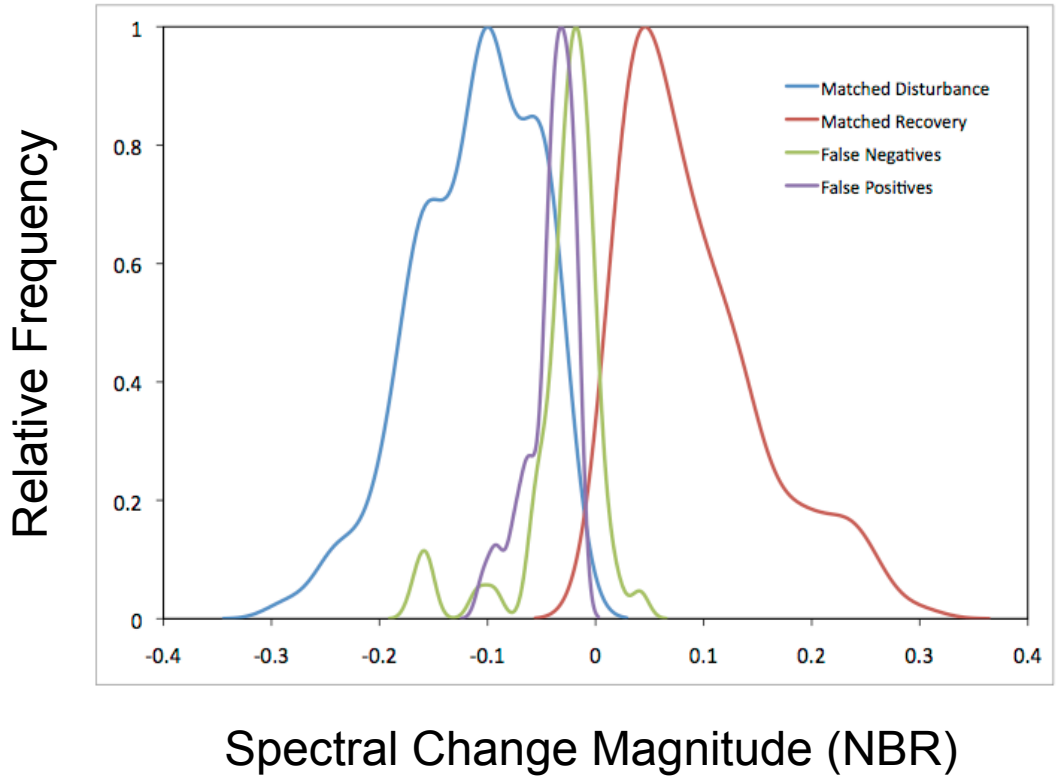
Relaxed Vertex Match Score Omission

Algorithm (NBR)

TimeSync	*Disturbance*	*Recovery*	*Stable*	*None*	*LT Omission*	*LT Omission by Type*	*LT Omission by Intensity*
Harvest - High	45	0	1	4	**0.10**		
Harvest - Medium	28	0	2	8	**0.26**		
Harvest - Low	14	3	4	17	**0.63**		
Harvest - All	87	3	7	29		**0.31**	
Fire - High	6	0	0	1	**0.14**		
Fire - Medium	17	0	0	2	**0.11**		
Fire - Low	5	0	0	0	**0.00**		
Fire - All	28	0	0	3		**0.10**	
Pathogen - Medium	4	0	0	2	**0.33**		
Pathogen - Low	4	0	0	4	**0.50**		
Pathogen - All	8	0	0	6		**0.43**	
Other - High	2	0	0	0	**0.00**		
Other - Low	1	0	0	2	**0.67**		
Other - All	3	0	0	2		**0.40**	
All - High	53	0	1	5			**0.10**
All - Medium	49	0	2	12			**0.22**
All- Low	24	3	4	23			**0.56**

- Big story here is disturbance intensity, with nearly all high and medium intensity disturbances captured, and nearly half of low intensity disturbances
- Low intensity disturbances nearly impossible to get with 3-year or greater interval
- Most medium intensity disturbances also missed with 3-year or more interval
- Also see that fires are perhaps most easily detected of all types

False Positive and Negative Disturbances (Relaxed)

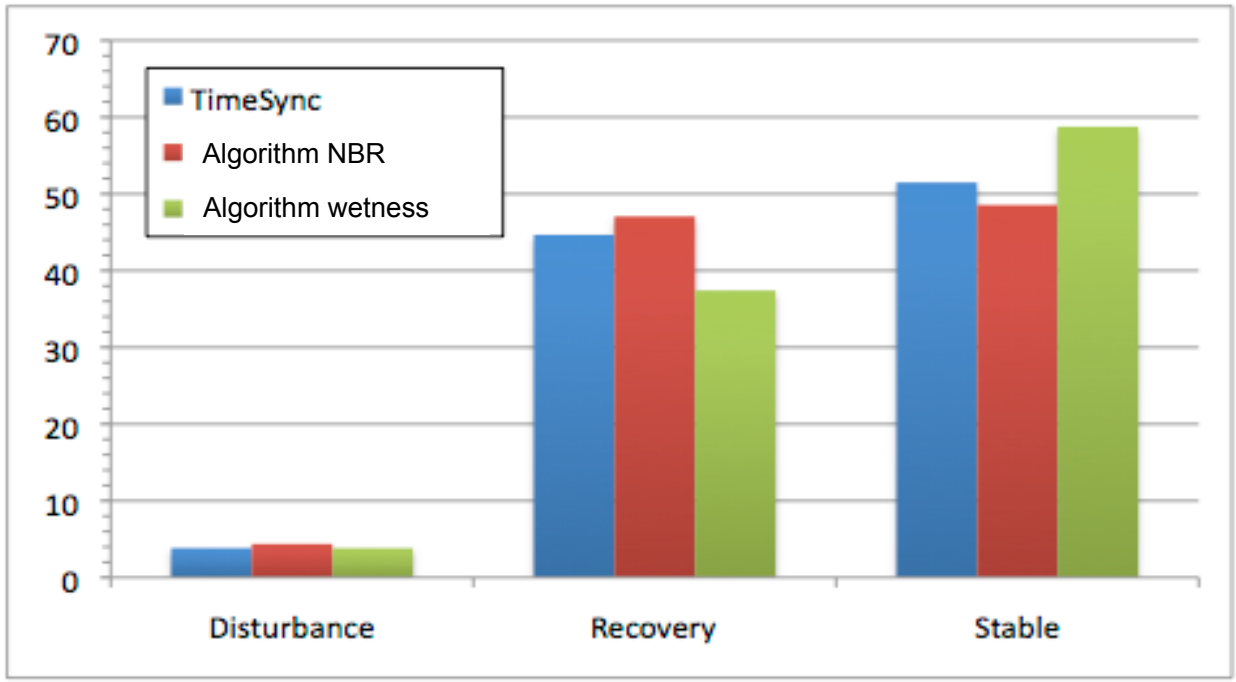


- False positives and false negatives are around the margin of spectral change magnitude
- False positives largely associated with cloud and shadow, phenology, residual misregistration – improvement possible
- False negatives subtle in context of temporal spectral variability – some improvement may be possible via more targeted image data selection (phenology and sun angle), but probably approaching limits of detectability via Landsat

TimeSync v. Algorithm Comparisons – Many Possibilities

- Segment summary distributions
  - Number of segments per plot & across plots
  - Number by category (disturbance, recovery, stable)
- Overall match scores
  - Category label match by year across trajectory, overall and by category
    - All three categories; Disturbance v. No Disturbance
  - Examine worst cases using TimeSync and algorithm trajectory plot
- Vertex match score matrices
  - Segment vertex match by category
    - Restricted and relaxed (fuzzy) precision
  - Examine all false negatives and positives (as above), and spectral change magnitude histograms for agreement and disagreement
- Proportion in each major category across plots and time

Average Percent of Time in Each Segment Category



- Most generalized measure (across plots and time series)
  - TimeSync and algorithm agree quite well, wetness more conservative
- In all comparisons we could compare output from different algorithms, not just different indices for same algorithm

Can We Trust TimeSync Interpretations?

- Having compared TimeSync v. Algorithm with a statistical valid sample: How good were TimeSync interpretations?
  - Intersect TimeSync plots with existing databases (e.g., in the US, USFS & BLM, MTBS, and FHM data/maps)
  - Derive agreement matrices
  - Explore/explain disagreements
- Caveats
  - Management agency databases only from portions of the land base
  - May themselves contain errors or be incomplete
  - MTBS is fire only
  - FHM data complicated to use due to imprecision of spatial location and polygon-based
  - Disturbance only

Results (from 4 TSAs)

	MTBS (Intensity)				
TimeSync	*Low*	*Medium*	*High*	*Not Disturbed*	*Agreement*
Harvest	1				n/a
Fire	8	10	4	5*	0.81
Not Disturbed				360	1
Agreement	0.89	1	1	0.99	0.98

* Below MTBS size threshold; confirmed as fire by alternate dataset

Results (from 4 TSAs)

MTBS (Intensity)

TimeSync	*Low*	*Medium*	*High*	*Not Disturbed*	*Agreement*
Harvest	1				n/a
Fire	8	10	4	5*	0.81
Not Disturbed				360	1
Agreement	0.89	1	1	0.99	0.98

* Below MTBS size threshold; confirmed as fire by alternate dataset

Integrated FACTS - FOI - FHM

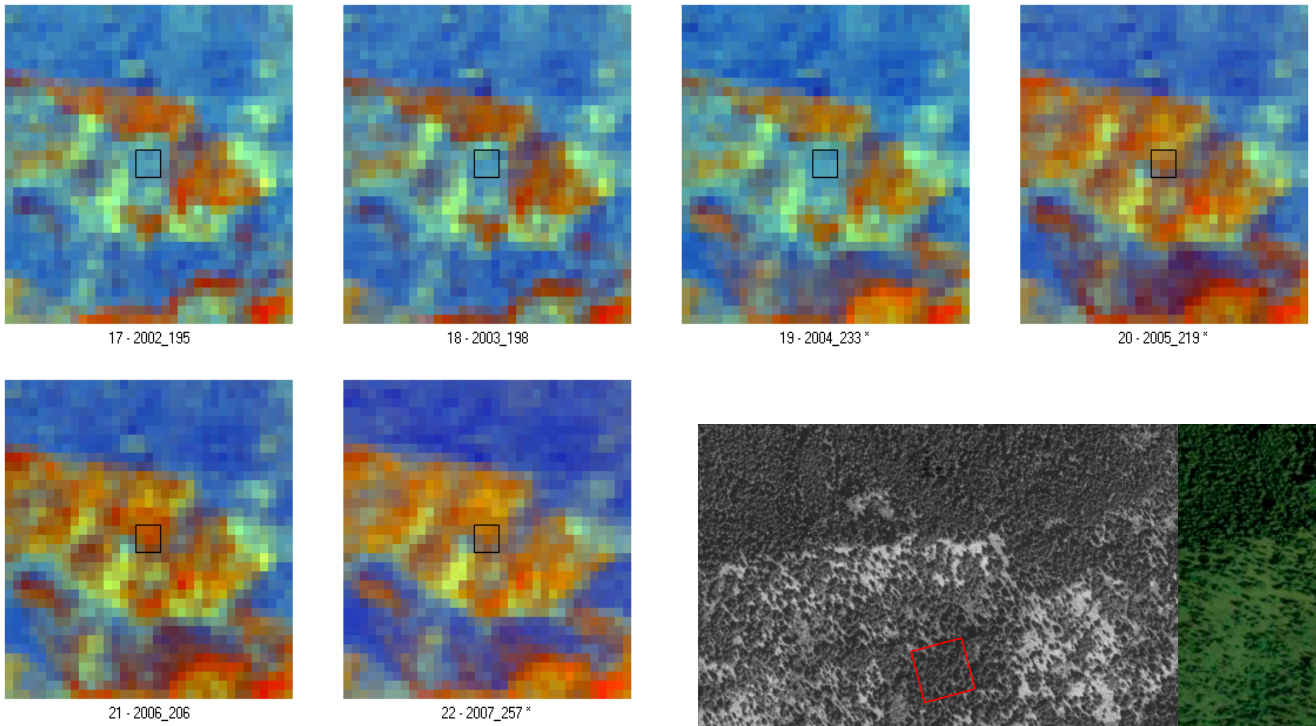
TimeSync	*Harvest (commercial)*	*Harvest (pre-commercial)*	*Understory Burn*	*Insect*	*Road*	*Not Disturbed*	*Agreement*
Harvest	8	2				10	0.5
Insect				7		1	0.88
Road						3	0
Not Disturbed	1	7	2			143	0.93
Agreement	0.89	0.22	0	1	0	0.91	0.87

TimeSync false negative

Agency database false negative

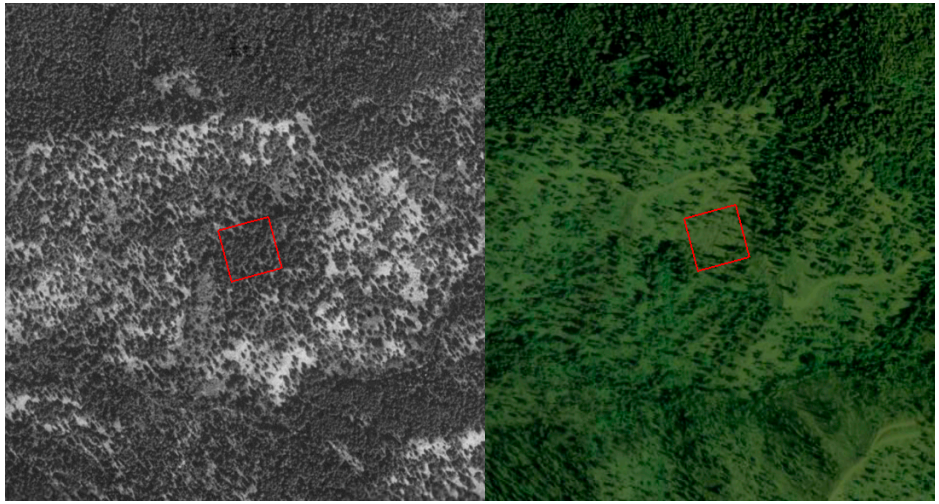
Agency database false negative

- Database missing harvest



Partial time-series

2006

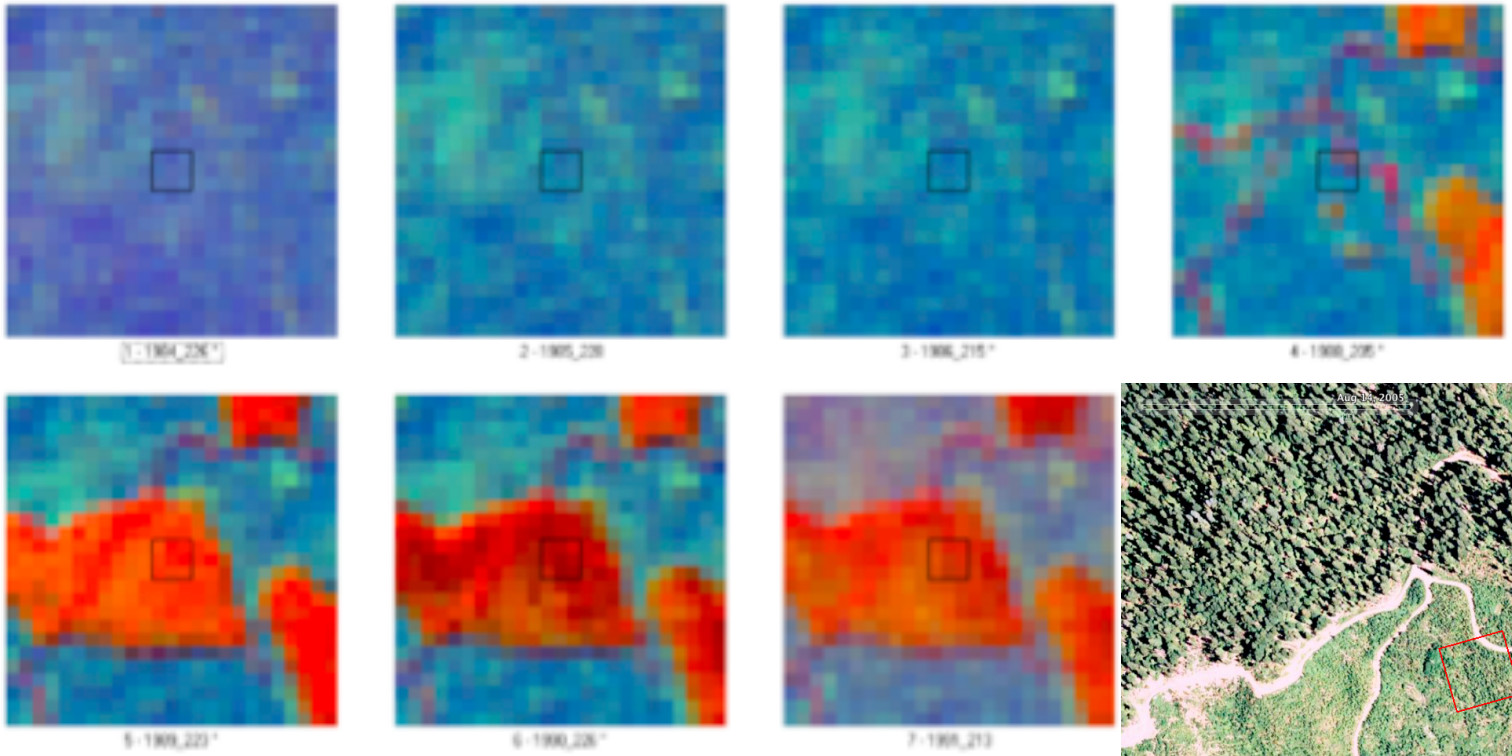


TimeSync plot: 45-27 #38

1998

Agency database false negative

- Database missing road



Partial time-series

TimeSync plot: 45-29 # 22

Summary

- We've entered a new frontier in Landsat change detection
  - More high quality data than ever before, thanks to USGS & Co.
  - New algorithms exploiting those data for change at annual time-step
  - New challenges for assessing map quality - lack of adequate reference data and methods
- TimeSync developed to address this need
  - Uses same time-series dataset as an automated algorithm, allowing for statistically robust plot sample designs
  - Based on old-fashioned airphoto-interpretation techniques
- Plots treated as trajectories having a series of labeled segments (disturbance, recovery, stable) with start date, duration, and spectral magnitude

Summary

- Demonstrated some new ways to evaluate annual change map quality
  - Segment summaries
  - Overall match scores
  - Vertex match scores (restricted and relaxed precision)
  - Average (across plot and time) proportion of time in disturbance, recovery, and stable
- Conducted a reality-check on TimeSync observations using existing agency disturbance databases
  - Learned what can't be easily detected with Landsat
  - Learned that even agency databases can't be trusted to represent reality

Conclusions & Notes

- TimeSync is an effective tool for assessment of change maps derived from Landsat time series algorithms
- We demonstrated some effective means of map assessment, but others are possible
- No one dataset or method contains the full truth; at best we can examine agreement amongst datasets and methods
- Stability, as used here, refers to spectral stability, not ecosystem stability
- With some additional improvements in data selection and processing we may be approaching the limits of change detection with Landsat data

Conclusions & Notes

- More explicit multivariate solutions may improve algorithm performance
- The bar has been raised
  - We now obtain a lot more information from Landsat time series
  - We obtain at least as accurate information as before with high intensity disturbances
  - We now obtain more accurate information for moderate intensity disturbances
  - Without annual time series, it is hardly possible to obtain information about low intensity disturbances
  - Our error/agreement matrices contain all of this information, so judging algorithm performance on simple overall agreement or Kappa statistics would be misleading