NDVI Data Reduction for a Fuzzy Statistical Evaluation

Jürgen Symanzik
Utah State University
Department of Mathematics and Statistics

Robert Gillies, Hee Lee, Peter Ma
Utah State University
Department of Aquatic, Watershed, and Earth Resources

Presentation Outline

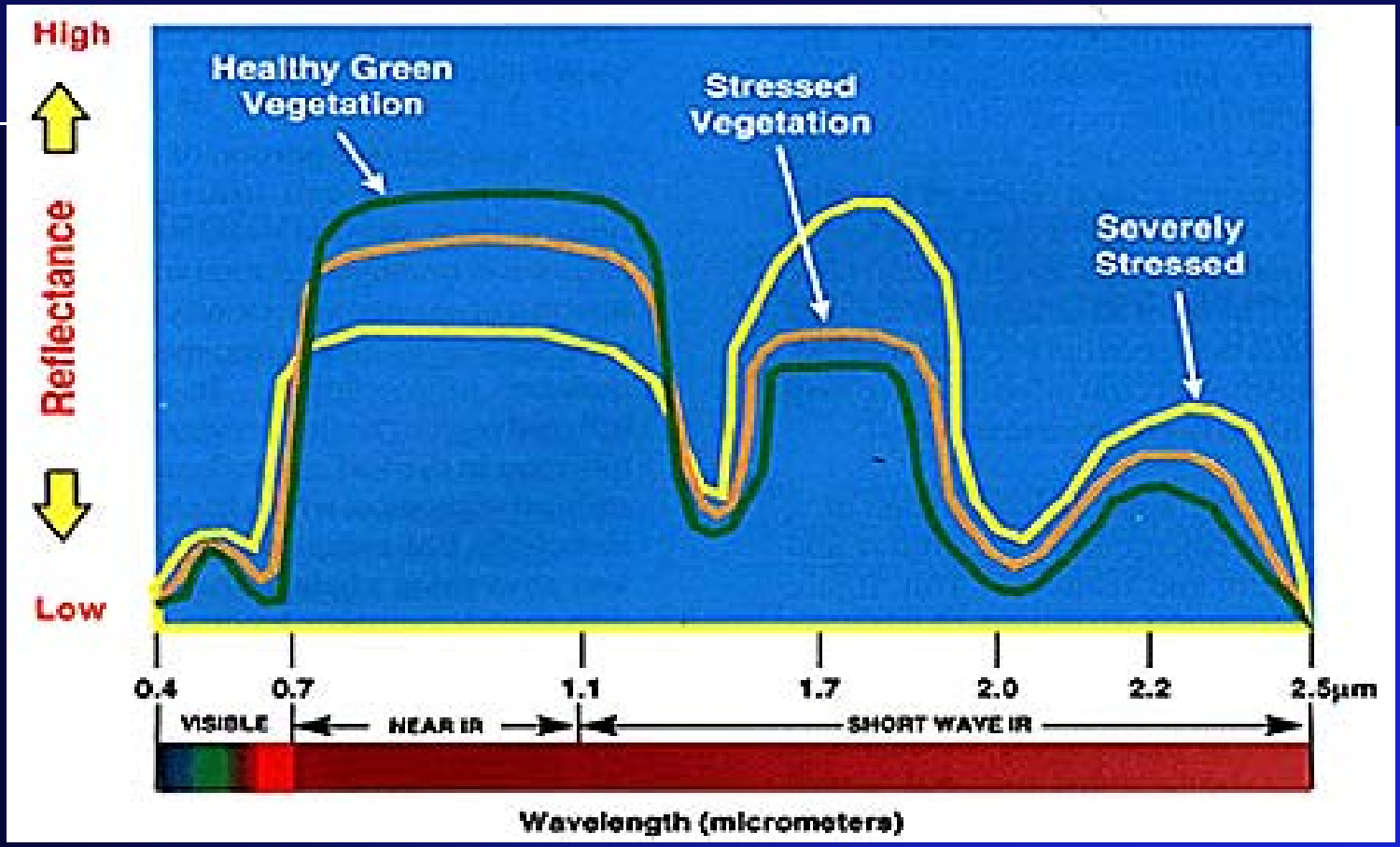
Background

- Remote sensing and vegetation

Development of NDVI Baseline Methodology

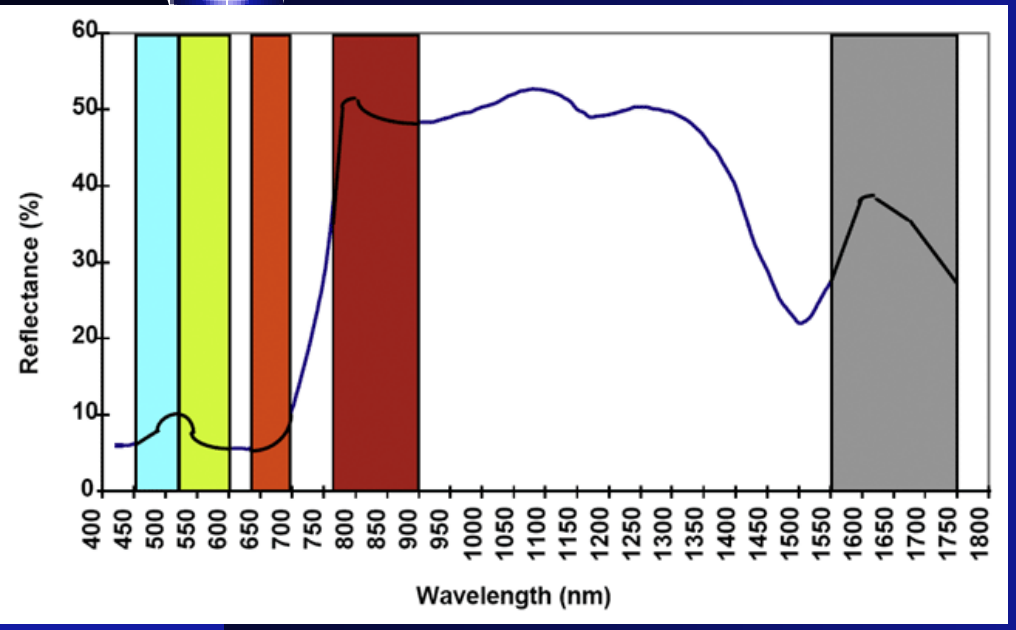
- Analysis of vegetation differences through probability density functions (PDFs) & cumulative distribution functions (CDFs)

Remote Sensing and Vegetation



- Energy reflectance measured by Earth Observation Satellites
- Monitoring "environmental" signature of vegetation

Normalized Difference Vegetation Index (NDVI)

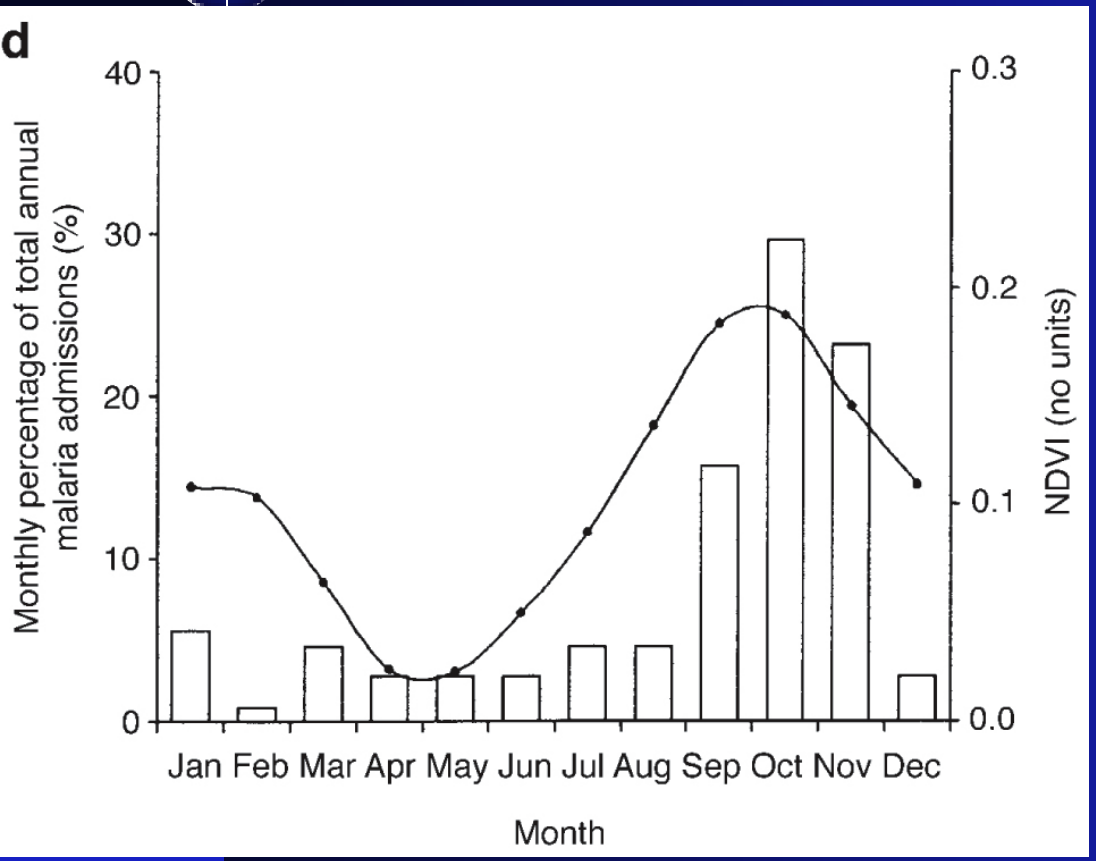


NDVI = (NIR–Red) / (NIR+Red)

Scale (-1 ≤ **NDVI** ≤ 1)

- Represents amount, type, and health of vegetation
- Shown to be associated with temperature & precipitation
- Related to vector-borne diseases

NDVI and Vector-Borne Diseases



- Commonly used

- Malaria: Hay 1998
- 30 day lag between peaks

- Malaria:
  Conner 1999
- Rift Valley Fever:
  Anyamba 2002
- West Nile Virus:
  Brownstein 2002

Questions of Interest for Epidemiologists

Is NDVI signature over a geographic region in year X different from NDVI signature in baseline ("normal") years, thus indicating a higher/lower risk for particular diseases?

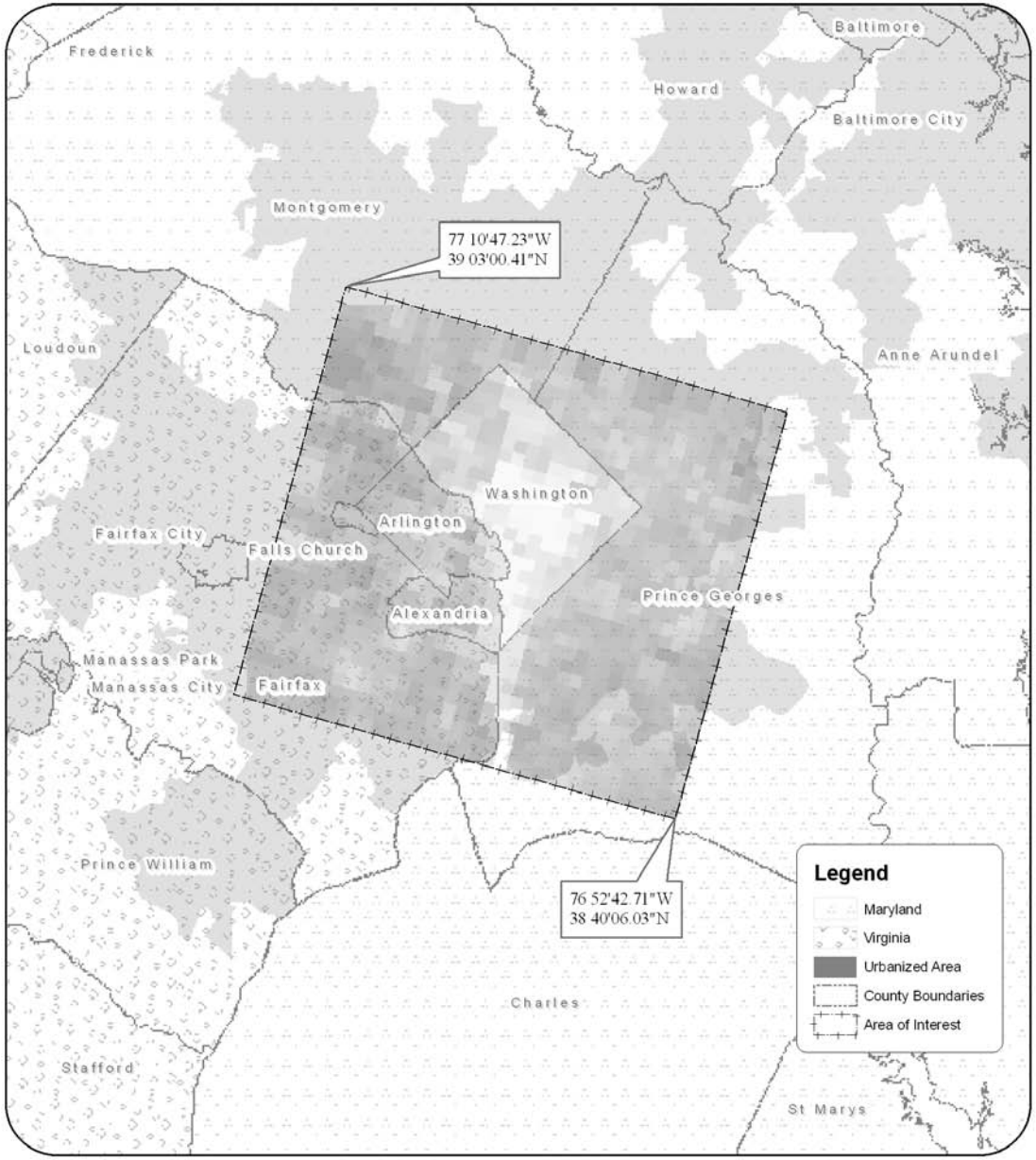
Problems with the data:

1. NDVI measurements at nearby locations not independent
2. NDVI distribution not always unimodal; outliers present
3. Comparison of means not satisfactory
4. Large number of NDVI measurements, making a significant departure from the baseline very likely
5. Not interested in reject/do not reject outcome of a single test, but rather a fuzzy classification needed: very similar, some departures, clearly different

Development of NDVI Baseline Methodology

Introduce a methodology to analyze NDVI using CDFs & PDFs

1. Define Study Area
2. Baseline Methodology
3. Introduce Statistical Methods
4. Results



Washington D.C. Area of Interest
AVHRR NDVI

0 1 2 Miles
0 2.5 5 Kilometers
N

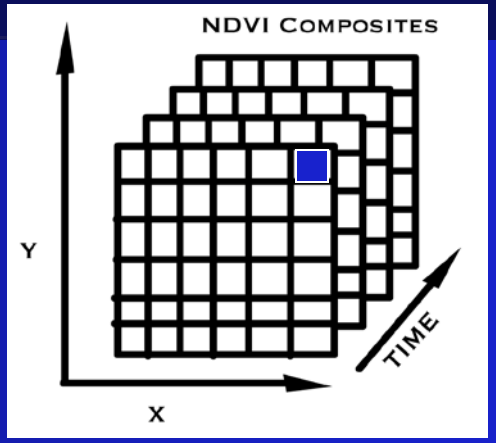
Date: 7/19/05
Geographic Coordinate System
by: Peter Ma

Study Area: Washington D.C.

- Strong seasonal NDVI phenology (climate/ forest)
- 135 Sq. km
- population of 570,000

METHODS: Baseline

- Crate a baseline, representing the typical (“normal”) NDVI signature
- Compare our “normal” against a particular year of interest (YOI)
- Use CDFs & PDFs to find significant differences



- Advanced Very High Resolution Radiometer (AVHRR)
- 14 years 1989-2003
- 1 km
- 26 composites per year
- 14-day intervals
- 35x33 pix = 1155pix

Two Baseline Methods

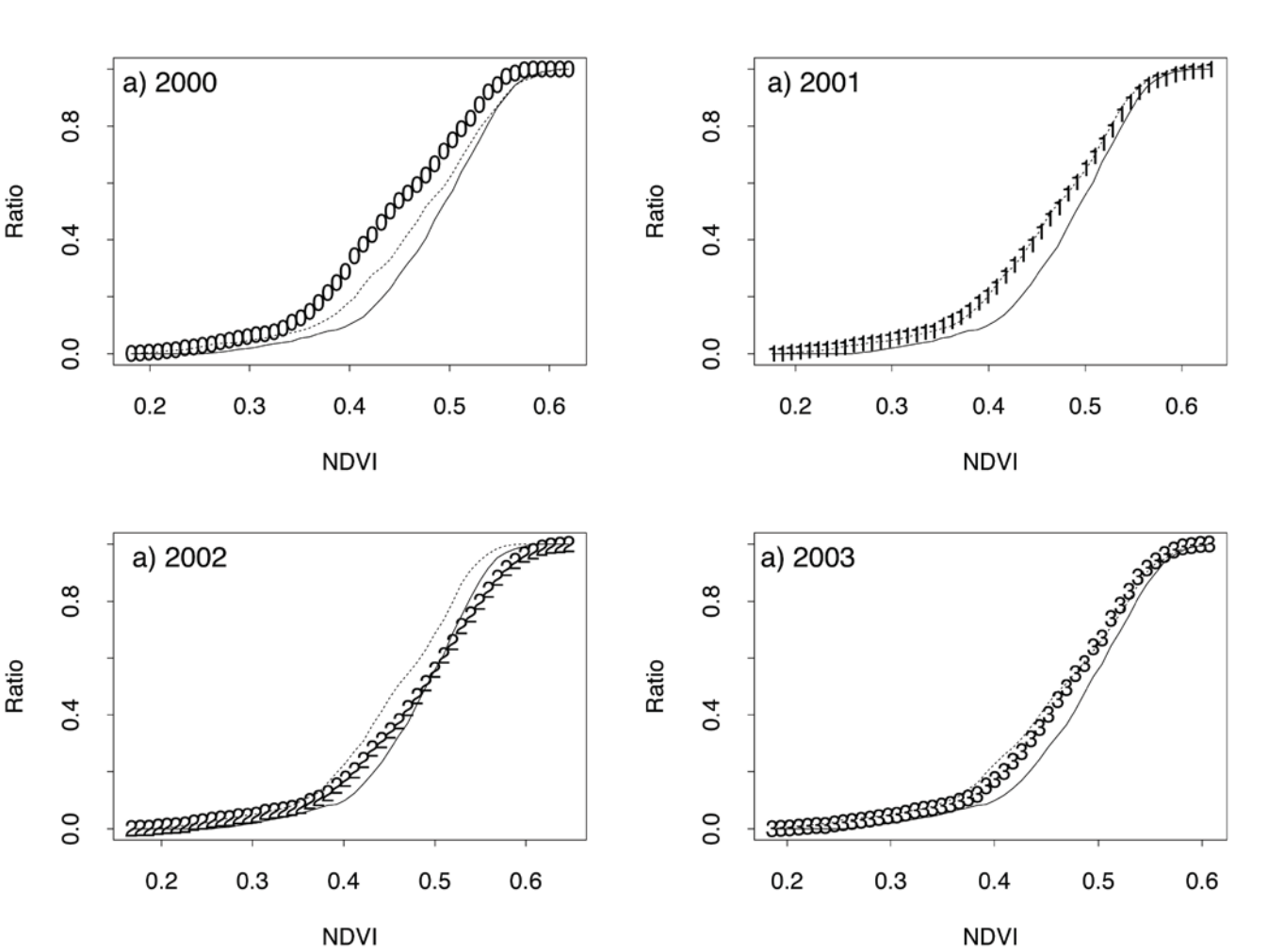
11-year

- "Long Term" – 11 years
- 1989 - 1999 => Average
- Historically rich timeline
- Commonly used in phenology studies

Comparative

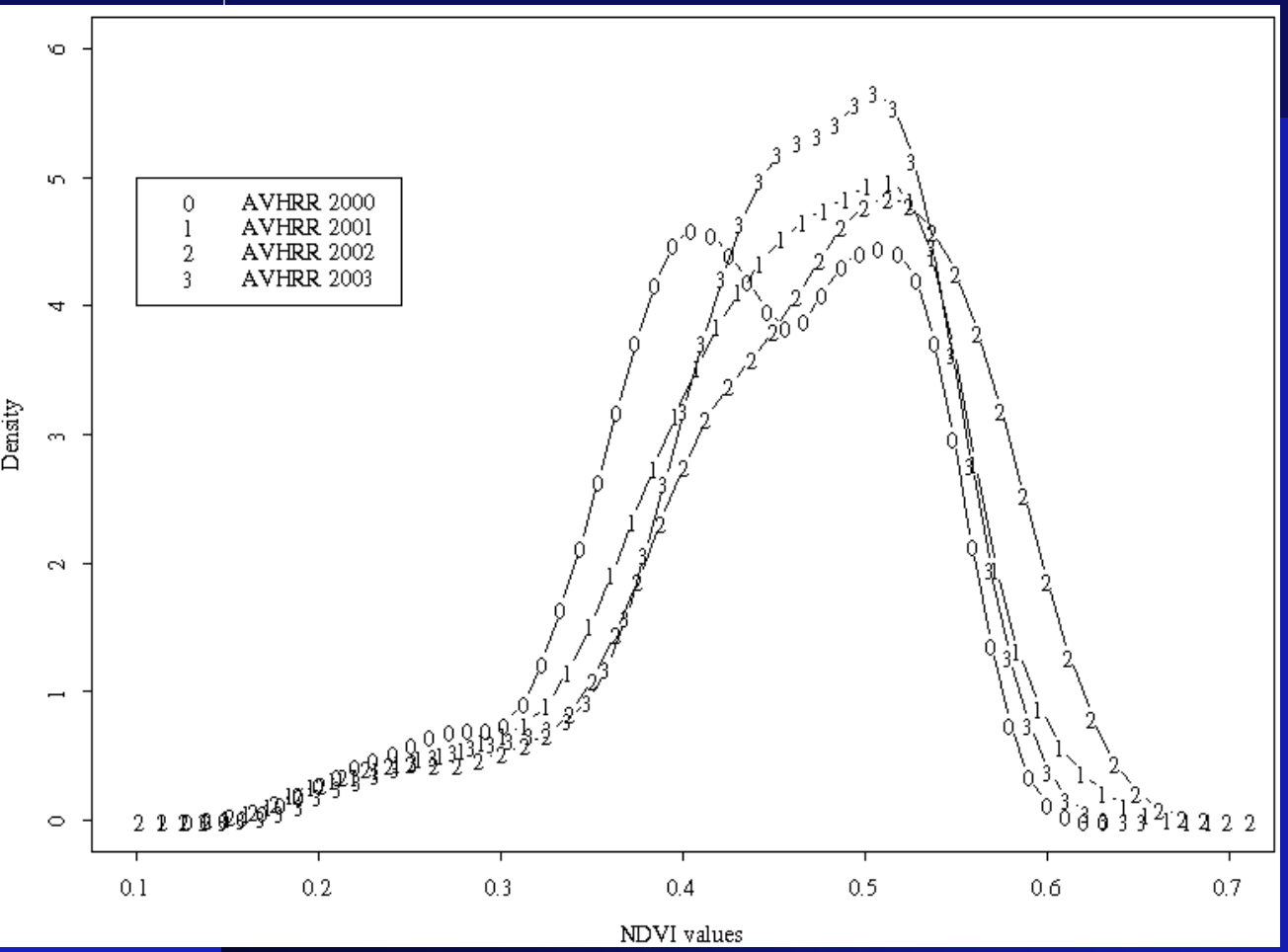
- "Short term" – 3 years
- Mean of surrounding years
- Comparative baseline for 2002 generated from mean of 2000, 2001, 2003 years

NDVI CDFs



YOI (0,1,2,3) vs Baselines (solid: 11-year, dashed: comparative)
Washington D.C.

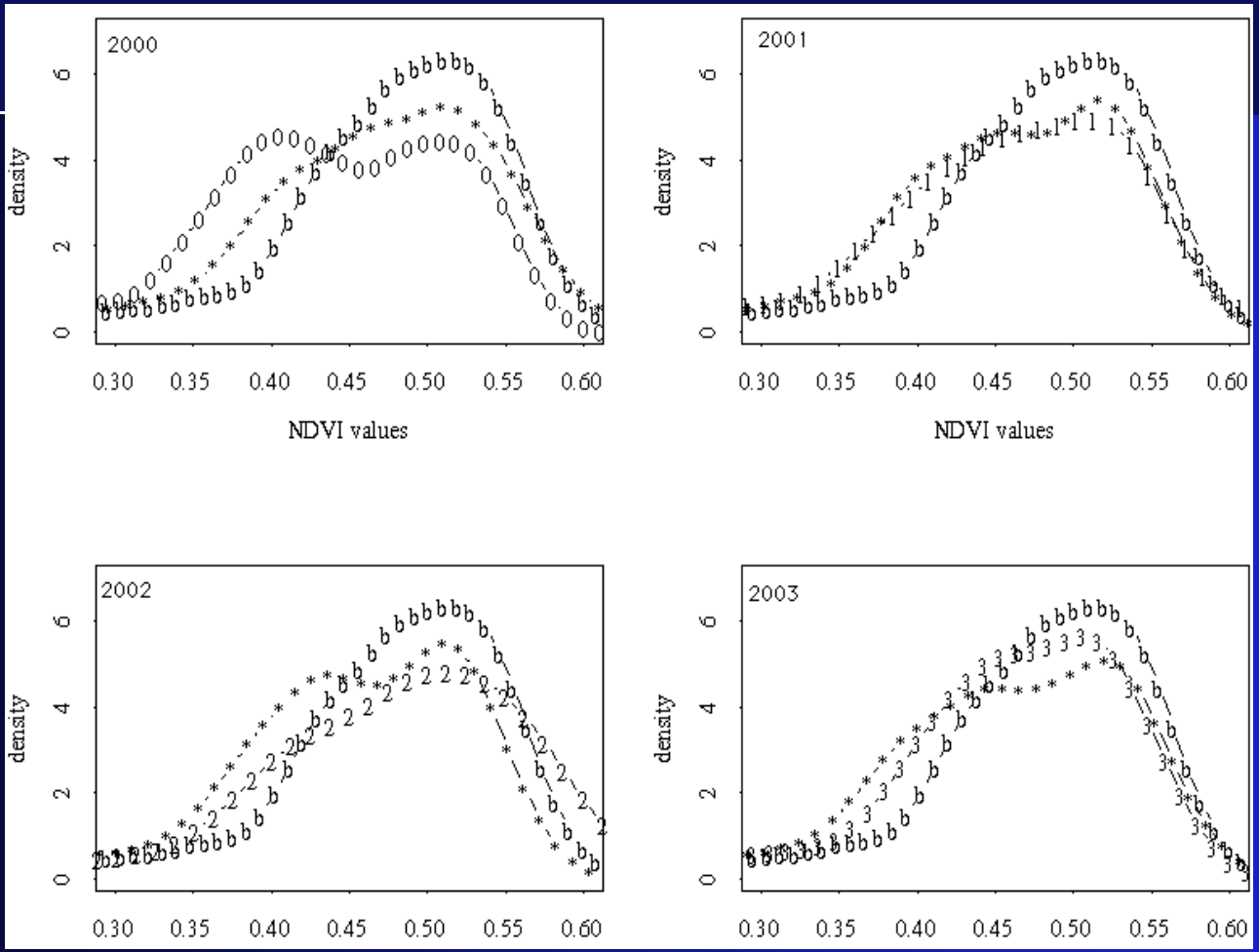
NDVI PDFs



Some shifts in NDVI distribution during these 4 years

Washington D.C. PDFs

NDVI PDFs



YOI (0,1,2,3) vs. Baselines (b: 11-year, *: comparative)
Washington D.C.

Tests (1)

Chi square / Kolmogorov-Smirnov Goodness of fit tests:

- Quantify differences observed in baseline and YOI PDFs and CDFs

Hypothesis Test

Null: Two populations have the same underlying distribution

Vs.

Alternative: Two populations do not have the same underlying distribution

Tests (2)

Chi square / Kolmogorov-Smirnov Goodness of fit tests:

Chi square ==> PDFs

K.S. ==> CDFs

- Both test for a distribution departure of YOI from baseline

Effects of Sample Size on Test Outcomes

- Thousands of pixels generate significant results for each test even for minor departures
- Recall: NDVI measurements are not independent
- Solution: Sample pixels from baseline and YOI, assuming sampled pixels will be independent
- Have to select appropriate sample size first

Evaluation of Sample Size for Normal Distribution

Dataset		N(0,1)A			N(0.5, 1)A			N(1, 1)A		
	Sample Size	50	100	200	50	100	200	50	100	200
N(0,1)B	50	**5.16** (4.07)			**40.23** (52.88)			**96.00** (96.87)		
	100		**4.91** (3.77)			**72.19** (82.29)			**99.97** (100.00)	
	200			**5.21** (5.04)			**85.91** (94.05)			**100.00** (100.00)
N(0.5, 1)B	50	(52.88)			**4.50** (3.73)					
	100		(82.29)			**4.63** (3.88)				
	200			(94.05)			**4.60** (3.51)			
N(1, 1)B	50	(96.87)						**4.55** (4.10)		
	100		(100.00)						**4.66** (3.77)	
	200			(100.00)						**4.90** (5.49)

- 2 generated random normal distributions (A,B)
- 3 sample sizes assessed (50, 100, 200)
- Same std (1) and three different means (0, 0.5, 1)
- # represent % of significant results from 10,000 iterations of Chi square (bold) and Kolmogorov-Smirnov (in parenthesis)
- Increased sample size = Increased % of significant results

Evaluation of Sample Size for NDVI Data

Sample Size	Statistics results from 25 Tests: Baseline[1] Chisq | Ks.gof				Statistics results from 25 Tests: Comparative Baseline[2] Chisq | Ks.gof			
	2000	2001	2002	2003	2000	2001	2002	2003
50	14 | 17	2 | 4	5 | 3	2 | 7	5 | 8	0 | 3	9 | 10	0 | 0
100	24 | 25	7 | 16	10 | 6	6 | 12	11 | 11	2 | 0	18 | 14	2 | 4
200	25 | 25	19 | 23	15 | 7	13 | 19	23 | 25	1 | 1	24 | 23	2 | 2

[1] Baseline = 11 years dataset (1989 to 1999)
[2] Comparative Baseline (e.g., 2002 was generated by 2000, 2001, and 2003 year dataset)
Chisq = Chi-square test
Ks.gof = Kolmogorov-Smirnov Goodness of Fit Test

- Verify affects of sample size observed in results of simulated normal distribution, using equally wide intervals (Standard Interval)
- Three sample sizes (50, 100, 200) assessed for NDVI data YOI and Baselines each
- 25 iterations of Chi square and Kolmogorov-Smirnov test

Test Iterations and the Binomial Distribution

Questions: How many significant results indicate a meaningful departure from the baseline?

For n = 25 tests:

Significant result recorded as P-value ≤ 0.05

Bin (25, 0.05) distribution:

No. of Success/25 trials	0	1	2	3	4	5	6	7	8
probability of success	2.77E-01	3.65E-01	2.31E-01	9.20E-02	2.67E-02	5.95E-03	1.04E-03	1.49E-04	1.77E-03
Sum of probability	0.28	0.64	0.87	0.97	0.99	1.00	1.00	1.00	1.00

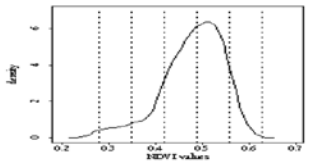
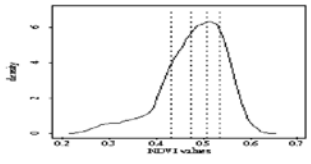
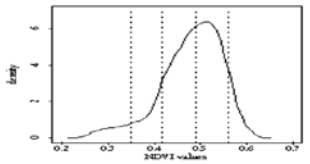
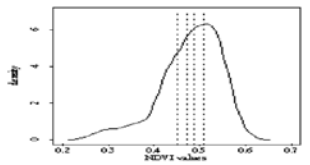
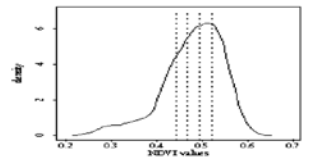
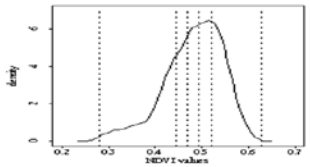
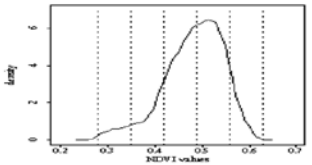
0 to 3: little evidence that distributions are different
4 to 9: some (to strong) evidence that distributions are different
10 or more: clear evidence that distributions are different

Categorization

1. Chi square and K.S. test used for **PDFs & CDFs**: Tests significant differences between baselines and YOI

2. Determined sample size of 50

3. Chi square requires continuous data (NDVI) to be categorized into discrete classes
   - Like sample size, categorization directly effects outcome of statistics
   - Developed / tested 7 methods to categorize NDVI baseline and YOI distributions

	Categorization Methods	Interval Range	% of Pixels	No. of Pixels	Density plots
a)	**Standard Interval** 1143 pixels (First sampled entire range then assigned intervals, resulted in sampled pixels from outside assigned interval)	0.28 ~ 0.35 0.35 ~ 0.42 0.42 ~ 0.49 0.49 ~ 0.56 0.56 ~ 0.63	4.29% 10.06% 34.91% 42.87% 7.87%	49 115 399 490 90	
b)	**20 Percentiles** 1155 pixels	-1.0 ~ 0.433 0.433 ~ 0.475 0.475 ~ 0.508 0.508 ~ 0.536 0.536 ~ 1.0	20.00% 20.00% 20.09% 19.91% 20.00%	231 231 232 230 231	
c)	**Standard Interval Extended Tail** 1155 pixels	-1.0 ~ 0.35 0.35 ~ 0.42 0.42 ~ 0.49 0.49 ~ 0.56 0.56 ~ 1.0	5.28% 9.96% 34.55% 42.42% 7.79%	61 115 399 490 90	
d)	**First Variation Mean & STD** mean ± 0.4STD and mean ± 0.1STD 1155 pixels	-1.0 ~ 0.455 0.455 ~ 0.475 0.475 ~ 0.488 0.488 ~ 0.508 0.508 ~ 1.0	30.30% 9.61% 8.92% 10.65% 40.52%	350 111 103 123 468	
e)	**Second Variation Mean & STD** mean ± 0.6STD and mean ± 0.2STD 1155 pixels	-1.0 ~ 0.442 0.442 ~ 0.468 0.468 ~ 0.494 0.494 ~ 0.521 0.521 ~ 1.0	23.55% 12.47% 16.62% 16.71% 30.65%	272 144 192 193 354	
f)	**Truncated Mean & STD** mean ± 0.6STD and mean ± 0.2STD 1143 pixels	0.28 ~ 0.446 0.446 ~ 0.471 0.471 ~ 0496 0.496 ~ 0.521 0.521 ~ 0.63	25.02% 11.64% 16.45% 16.01% 30.88%	286 133 188 183 353	
g)	**Truncated Standard Interval** 1143 pixels (First truncated range then sampled)	0.28 ~ 0.35 0.35 ~ 0.42 0.42 ~ 0.49 0.49 ~ 0.56 0.56 ~ 0.63	4.29% 10.06% 34.91% 42.87% 7.87%	49 115 399 490 90	

- First: sampled
- 98% pixels (0.28-0.63)
- standard interval, width 0.07

- First: sampled
- Intervals, each contain 20%
- range -1 to 1

- First: sampled
- same as standard interval
- extended tails -1.0 to 1.0

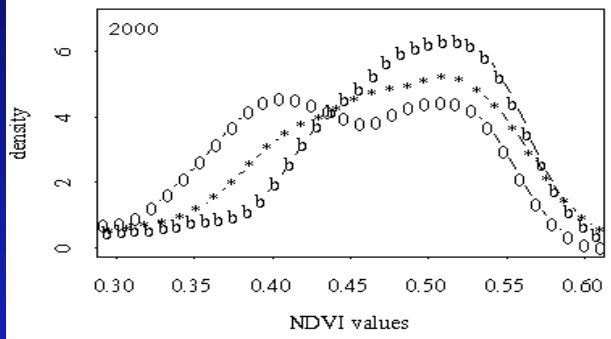
- First: sampled
- 11-year baseline
- mean &std: 0.4813 & 0.0656
- (0.4*std) (0.1*std) : 1st Var
- (0.6*std) (0.2*std) : 2nd Var

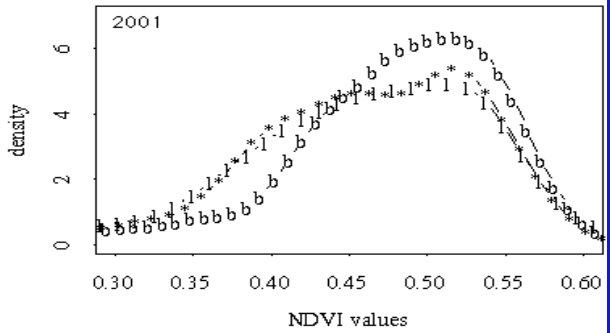
- First: truncated
- 98% pixels (0.28-0.63)
- Second: sampled
- mean & std: 0.4835 & 0.0625

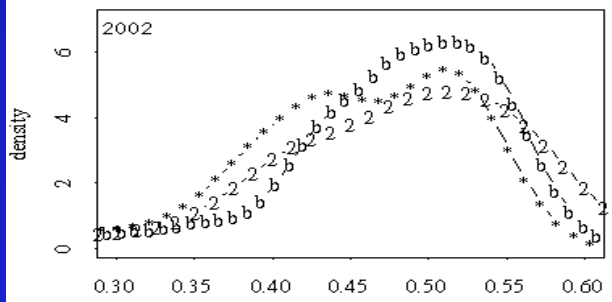
- First: truncated
- 98% pixels (0.28-0.63)
- standard interval, width 0.07

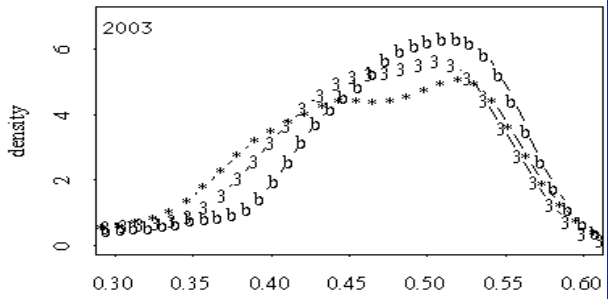
	Categorization Methods (50 Random Samples)	Significant Results from 25 Tests: Baseline[1)] Chisq | Ks.gof				Significant Results from 25 Tests: Comparative Baseline[2)] Chisq | Ks.gof			
		2000	2001	2002	2003	2000	2001	2002	2003
a)	**StandardInterval**	14 | 17	2 | 4	5 | 3	2 | 7	5 | 8	0 | 3	9 | 10	0 | 0
b)	**20 Percentiles**	15 | 17	4 | 4	4 | 3	6 | 7	6 | 8	3 | 3	10 | 10	0 | 0
c)	**Standard Intervals Extended Tail**	14 | 17	3 | 4	7 | 3	4 | 7	5 | 8	2 | 3	11 | 10	1 | 0
d)	**First Variation Mean & STD** mean ± 0.4STD and mean ± 0.1STD	11 | 17	2 | 4	0 | 3	5 | 7	9 | 8	0 | 3	5 | 10	0 | 0
e)	**Second Variation Mean & STD** mean ± 0.6STD and mean ± 0.2STD	14 | 17	4 | 4	3 | 3	6 | 7	7 | 8	2 | 3	7 | 10	0 | 0
f)	**Truncated Mean & STD** mean ± 0.6STD and mean ± 0.2STD	13 | 17	4 | 6	2 | 3	2 | 4	4 | 7	0 | 2	3 | 4	1 | 0
g)	**Truncated Standard Interval**	19 | 17	7 | 6	6 | 3	3 | 4	7 | 7	1 | 2	6 | 4	1 | 0

[2)] Comparative Baseline (e.g., 2002 was generated by 2000, 2001, and 2003 year dataset)
Chisq = Chi-square test
Ks.gof = Kolmogorov-Smirnov Goodness of Fit Test









YOI (0,1,2,3)
vs.
Baselines

b: 11-year
*: comparative

Methods Conclusions

NDVI PDF / CDF Methods

Results of sample size and categorization analysis showed that sample size of 50 and *Second Variation Mean and STD* method are most appropriate

Sample Size : 50

- Observations from generated random normal distribution
- Results from NDVI analysis of sample sample size

Categorization: mean & std

- No data lost due to truncation
- Method is data driven
- # of significant test outcomes are closest to the median of all the results presented

Baseline Conclusions

11-year (long term) Baseline

2002: little departure

2001 & 2003: some departures

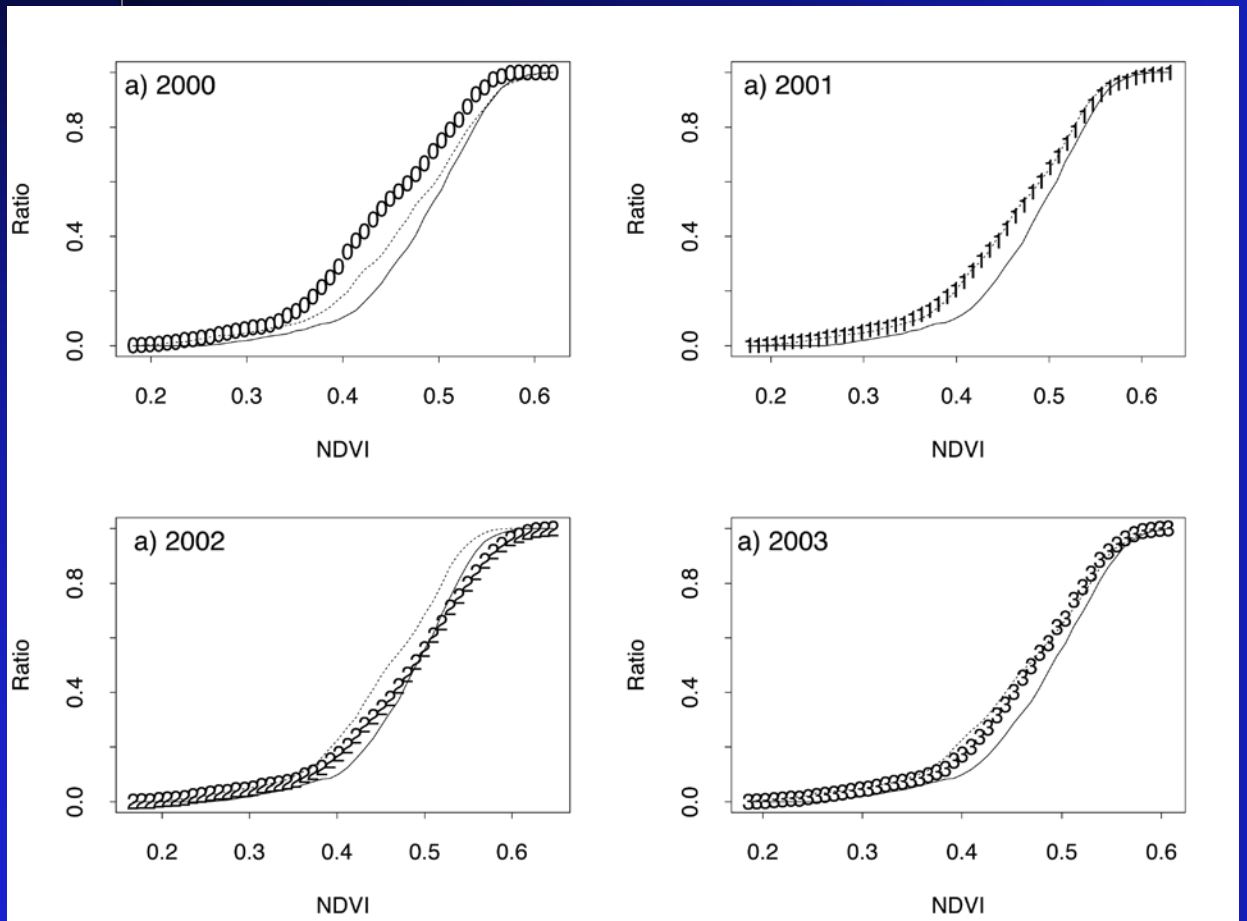
2000: clear departure

Comparative (short term) Baseline

2003 & 2001: little departures

2000: some departure

2002: some/clear departure



YOI (0,1,2,3)
vs.
Baselines

solid: 11-year
dashed: comparative

Questions ?