Best Practices for Classification Accuracy Metrics

Dr. Robert Gilmore Pontius Jr
rpontius@clarku.edu
www.clarku.edu/~rpontius

CLARK UNIVERSITY GRADUATE SCHOOL OF GEOGRAPHY EST. 1921

NASA Land-Cover and Land-Use Change Program

NATIONAL SCIENCE FOUNDATION LTER NETWORK LONG TERM ECOLOGICAL RESEARCH

MAPBIOMAS

Pontius' recommendations for Best Practices

1. Select a metric that addresses your research question, which is difficult.
2. Think in terms of quantity and allocation differences, which are concepts that popular metrics fail to distinguish.
3. Use the book Metrics That Make a Difference: How to Analyze Change and Error starting with the chapter *Commandments to Avoid Deadly Sins*.
4. Consider your motivations, which might conform to a flawed culture that reports accuracy without reporting the reference data's unreliability.
5. Get free materials at Pontius' website www.clarku.edu/~rpontius
6. Advise predoctoral colleagues to enter university programs, e.g. Clark University.
7. Discuss your problems openly to maximize learning.

Which of the Comparison Maps agrees more with the reference map?

Reference

1	0	0	0	0
0	0	0	0	0

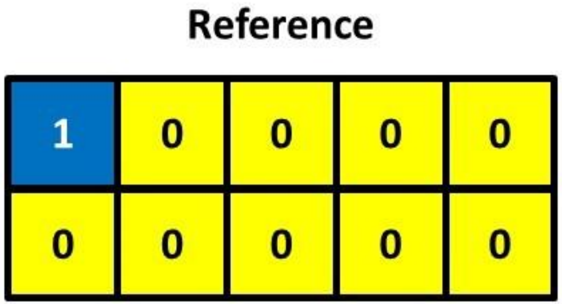
Comparison 1

1	0	1	1	1
1	1	1	1	1

Comparison 2

0	1	0	0	0
0	0	0	0	0

Which of the Comparison Maps agrees more with the reference map?



Multiple Choice
Comparison 1
Comparison 2
Other

Which of the Comparison Maps agrees more with the reference map?

Reference

1	0	0	0	0
0	0	0	0	0

Comparison 1

1	0	1	1	1
1	1	1	1	1

Comparison 2

0	1	0	0	0
0	0	0	0	0

Pontius selects Other because *agrees more* is insufficiently precise.

Pontius does not like the question because it focuses on agreement.
We are likely to learn more from difference than from agreement.

Which of the Comparison Maps agrees more with the reference map?

If *agrees* means number of matching pixels, then Comparison 2 agrees more than Comparison 1.

Many authors want to use an index on the range from 0 to 1 where 1 means perfect agreement and zero means something else.
Many authors want to report a number between 0.85 and 0.95.

[Wikipedia](https://en.wikipedia.org) has 20 indices for this situation of two classes.
The most popular index is percent correct.

Reference

1	0	0	0	0
0	0	0	0	0

Comparison 1

1	0	1	1	1
1	1	1	1	1

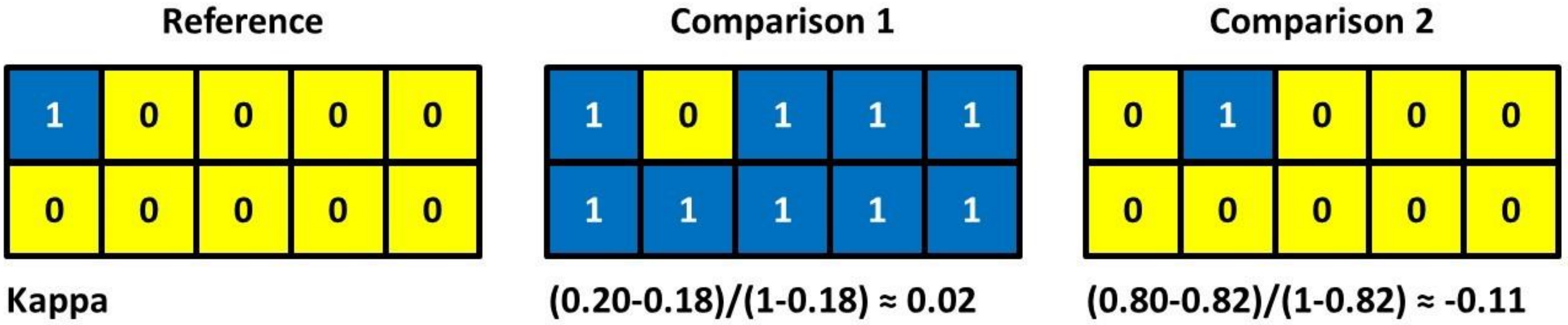
Comparison 2

0	1	0	0	0
0	0	0	0	0

Percent Correct says Comparison 2 agrees more than Comparison 1.
Percent Correct says that an all yellow map agrees more than Comparison 2.
It is dangerous to maximize a metric that you do not understand properly.

Pontius and Millones (2011) Death to Kappa. International Journal of Remote Sensing.
https://www.tandfonline.com/doi/abs/10.1080/01431161.2011.552923

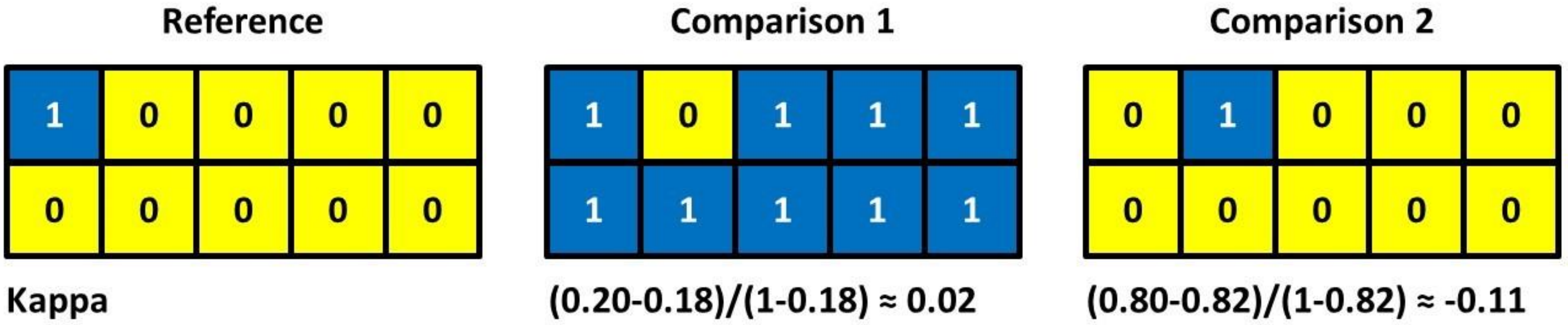
A popular metric is Kappa.



Kappa says Comparison 1 agrees more than Comparison 2.
If you select a metric based on its popularity, then you are likely to do absurd things.

Pontius and Millones (2011) Death to Kappa. International Journal of Remote Sensing.
https://www.tandfonline.com/doi/abs/10.1080/01431161.2011.552923

You must align your metric with your research question. You are likely to realize that you have a vague research question, in which case you have learned something.



Any universal rule for selection of a particular metric and the value of the metric for acceptability is absurd because any universal rule is not connected to any particular research question.

Anderson’s recommendation that percent correct should be greater than 85% is absurd and has caused horrendous damage to the profession.

Focus on the reasons for the disagreement.
Comparison 1 has a disagreement in quantity.
Comparison 2 has a disagreement in allocation.

Reference

1	0	0	0	0
0	0	0	0	0

Comparison 1

1	0	1	1	1
1	1	1	1	1

Comparison 2

0	1	0	0	0
0	0	0	0	0

	Comparison 1	Comparison 2
Quantity Disagreement	**0.80**	**0.00**
Allocation Disagreement	**0.00**	**0.20**

Pontius and Millones (2011) Death to Kappa. International Journal of Remote Sensing.
https://www.tandfonline.com/doi/abs/10.1080/01431161.2011.552923

If your purpose is to estimate the quantity, then comparison 2 is perfect.

Reference

1	0	0	0	0
0	0	0	0	0

Comparison 1

1	0	1	1	1
1	1	1	1	1

Comparison 2

0	1	0	0	0
0	0	0	0	0

	Comparison 1	Comparison 2
Quantity Disagreement	0.80	0.00
Allocation Disagreement	0.00	0.20

Pontius (2000) endorsed various forms of kappa.
Then Pontius realized his flawed thought process.

Pontius and Millones (2011) published the Death To Kappa, which had two messages:
Don’t use Kappa.
Use quantity and allocation disagreement.
The Death to Kappa paper has more than 1900 citations.

Pontius and Millones (2011) Death to Kappa. International Journal of Remote Sensing.
https://www.tandfonline.com/doi/abs/10.1080/01431161.2011.552923

Our literature review shows that half of the papers that cited the Death to Kappa paper still used Kappa.

Many papers that reported quantity and allocation difference failed to interpret the difference in a manner that relates to any research question. Many papers reported the metrics then concluded the results are acceptable without defining *acceptable*.

Pontius has not seen the use of the word *acceptable* applied in an intelligent manner for a practical question in his profession.

Pontius and Millones (2011) Death to Kappa. International Journal of Remote Sensing.
https://www.tandfonline.com/doi/abs/10.1080/01431161.2011.552923

Here is how some authors cite the *Death To Kappa* paper by Pontius and Millones (2011)

“kappa coefficient … has proved to be an excellent statistical parameter for measuring consistency (Pontius and Millones 2011).”

cited in Gao et al. (2021) https://doi.org/10.1016/j.ijdrr.2020.101928

If you want to compute agreement for a continuous variable, then consider this question.

What is the agreement between 5 and 2?

If you want to compute agreement for a continuous variable, then consider this question.

What is the agreement between 5 and 2?

The question is flawed because it lacks a definition of agreement.

If you want to compute difference for a continuous variable, then consider this question.

What is the difference between 5 and 2?

<u>Multiple Choice</u>
3
Other

If you want to compute difference for a continuous variable, then consider this question.

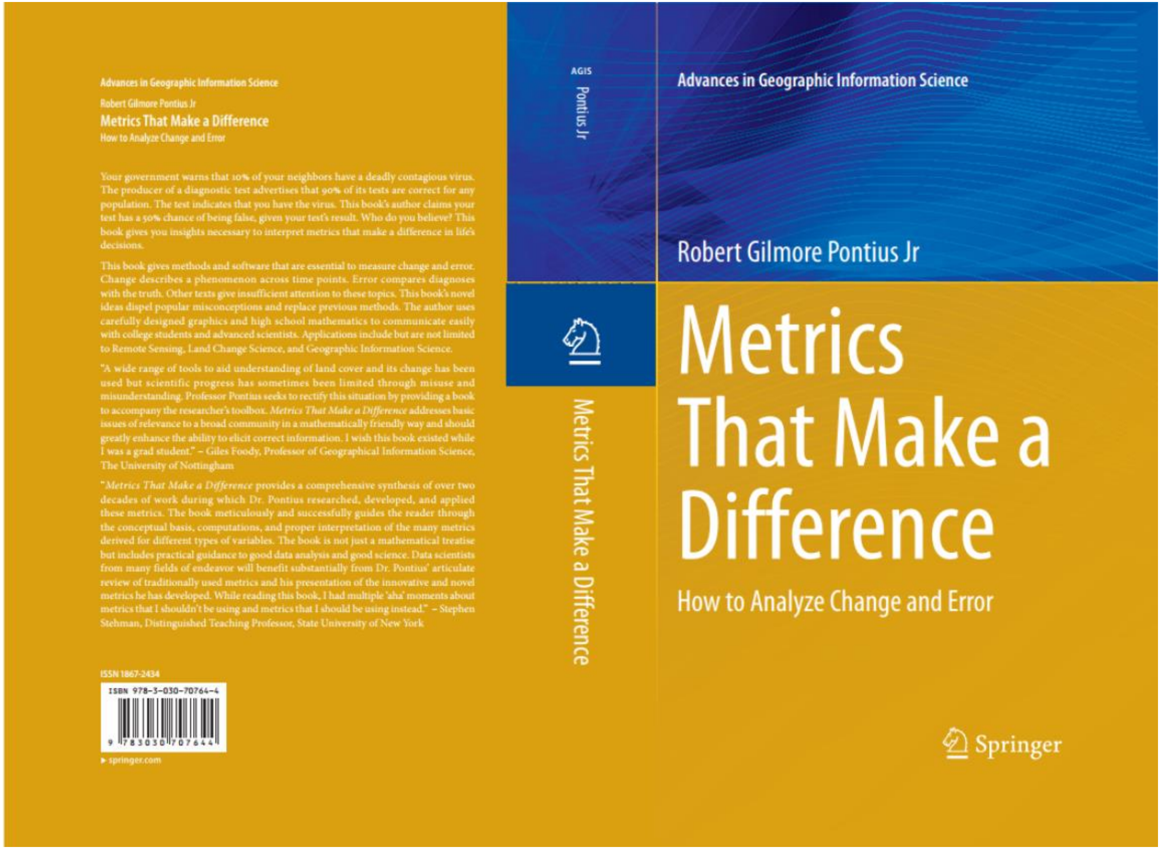
What is the difference between 5 and 2?

Pontius says Other because the definition of difference is vague.
The difference could be 5-2 = 3 or 2-5 = -3.

This exercise is helpful to refine the research question.
Think in terms of difference of quantity and allocation, which you can learn in the book on the following slide.

Ask your librarian to get this book

https://link.springer.com/book/10.1007/978-3-030-70765-1

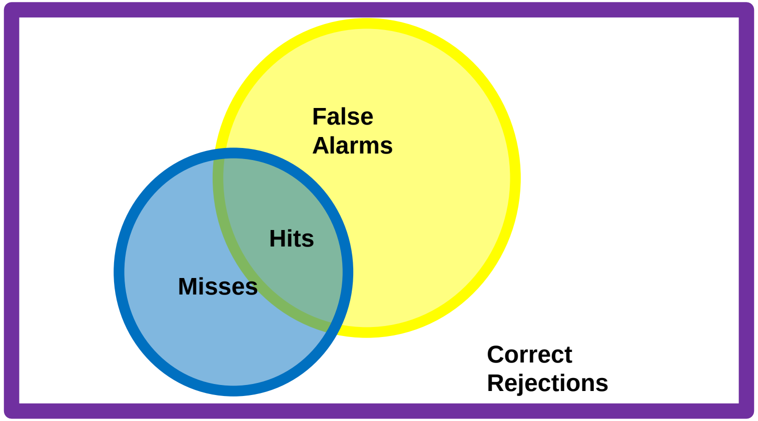


The book explains metrics for four common cases in chapters 1, 2, 4, and 8. You should start with Chapter 12.

Chapter 1

X	P	P	P	A	A	A	A	A	A	A
Y	P	A	A	P	P	P	A	A	A	A
Y versus X	H	F	F	M	M	M	C	C	C	C

		Y	
		Presence	**Absence**
X	**Presence**	Hits =1	False Alarms = 2
	Absence	Misses = 3	Correct Rejections = 4



The table is a rectangular Venn Diagram.

If Misses ≠ False Alarms, then Quantity disagreement is positive.

If Misses > 0 and False Alarms > 0, then Allocation disagreement is positive.

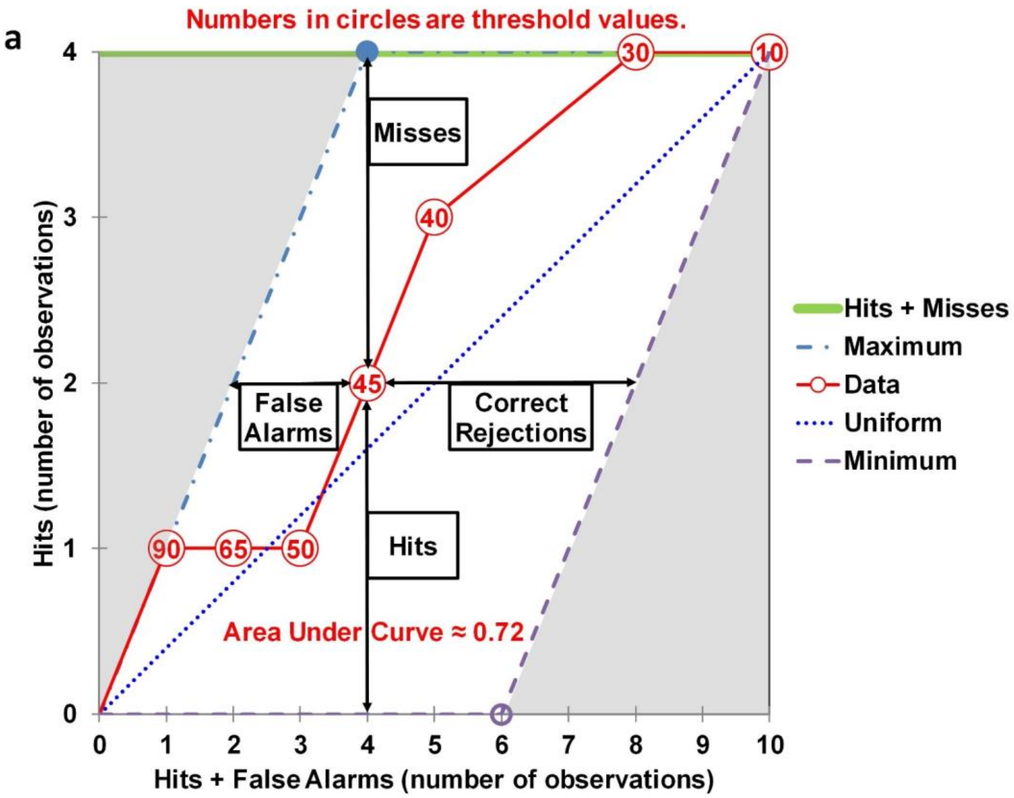
Chapter 2

X is indicates rank, not magnitude.
90 is ranked first
65 is ranked second
50 is ranked third.

X	90	65	50	45	40	30	30	30	10	10
Y	P	A	A	P	P	P	A	A	A	A

The Total Operating Characteristic (TOC) shows the values of all the entries in the contingency table at each threshold.

The TOC is more enlightening than the popular Relative Operating Characteristic (ROC).



Chapter 4

If you want to play with fire, then use more than two categories.

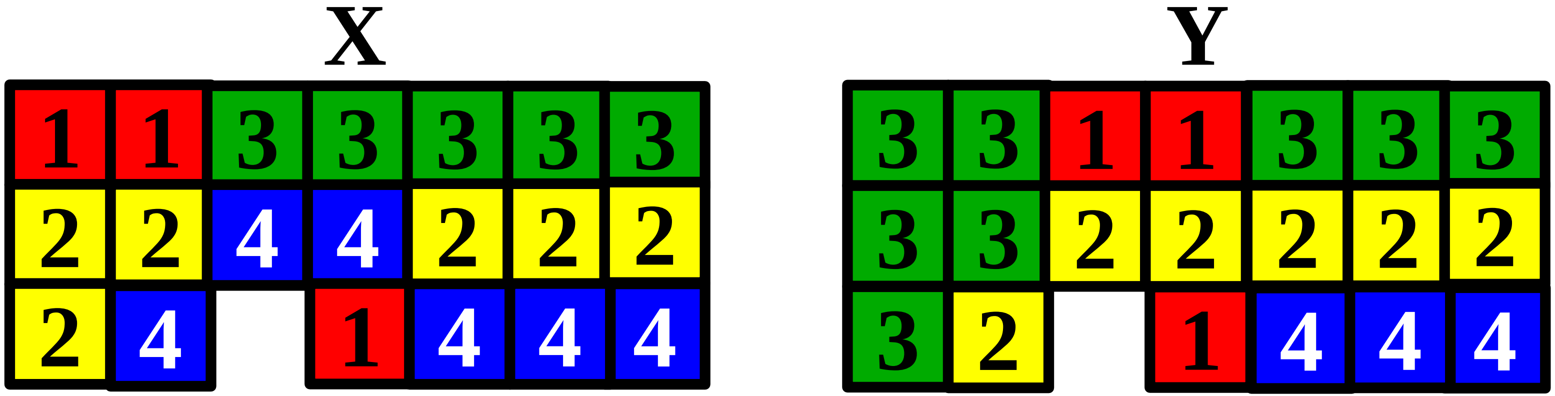
X and Y are two realizations of the same categorical variable.

Case 1: X is the classification, Y is the reference.

Case 2: X is an initial time, Y is a subsequent time.

Case 3: X is one classification, Y is another classification.

Use the concepts from Chapter 1 to make a table to think in terms of quantity and allocation.



Venn Diagram for category 1

		Y					
		j=1	*j*=2	*j*=3	*j*=4	**Sum**	**False Alarms**
X	*i*=1						
	i=2						
	i=3						
	i=4						
	Sum						
	Misses						

Venn Diagram for category 2

		Y					
		j=1	j=2	j=3	j=4	Sum	False Alarms
X	i=1						
	i=2						
	i=3						
	i=4						
	Sum						
	Misses						

Venn Diagram for category 3

		Y					
		j=1	*j*=2	*j*=3	*j*=4	**Sum**	**False Alarms**
X	*i*=1						
	i=2						
	i=3						
	i=4						
	Sum						
	Misses						

Venn Diagram for category 4

		Y					
		***j*=1**	***j*=2**	***j*=3**	***j*=4**	**Sum**	**False Alarms**
X	***i*=1**						
	***i*=2**						
	***i*=3**						
	***i*=4**						
	Sum						
	Misses						

With more than two categories, there are three components of difference: Quantity, Exchange and Shift

		Y					
		j=1	*j*=2	*j*=3	*j*=4	Sum	False Alarms
X	*i*=1						
	i=2						
	i=3						
	i=4						
	Sum						
	Misses						

		Miss				False Alarms		
		Quantity	Exchange	Shift	Hits	Shift	Exchange	Quantity
Category	1							
	2							
	3							
	4							

With more than two categories, there are three components of difference: Quantity, Exchange and Shift

Quantity indicates the size of each class.
Exchange indicates classes that are confused with each other.
Shift can show a pattern where Forest changes to Agriculture in some locations while Agriculture changes to Urban in other locations.

		Miss				False Alarms		
		Quantity	**Exchange**	**Shift**	**Hits**	**Shift**	**Exchange**	**Quantity**
Category	**1**							
	2							
	3							
	4							

Chapter 8 Interval versus Interval Variable

First step is to make at plot with identical axes and the Y=X diagonal line, then look at it!

https://www.autodeskresearch.com/publications/samestats

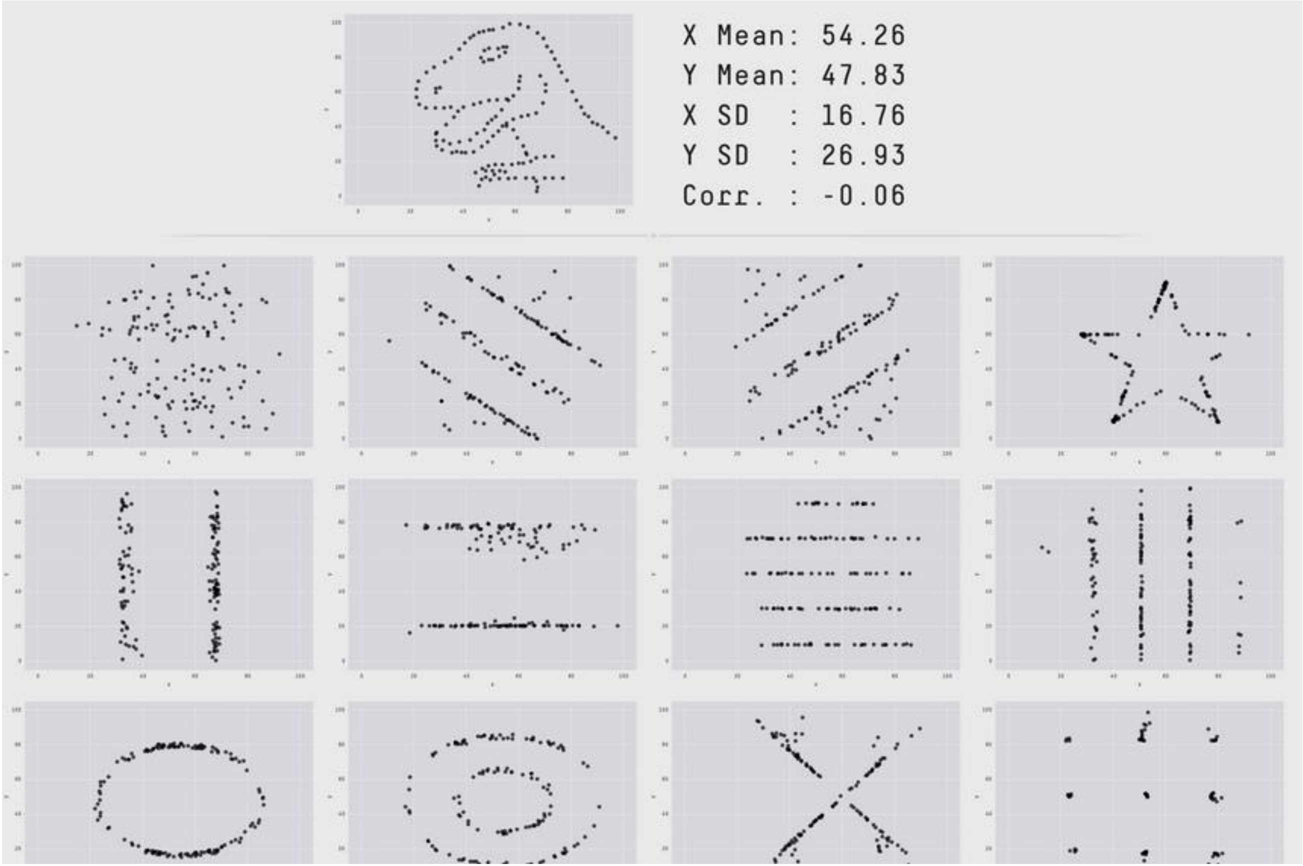
The Datasaurus Dozen

Recently, Alberto Cairo created the Datasaurus dataset which urges people to "never trust summary statistics alone; always visualize your data", since, while the data exhibits normal seeming statistics, plotting the data reveals a picture of a dinosaur. Inspired by Anscombe's Quartet and the Datasaurus, we present, The Datasaurus Dozen (download .csv):

These 13 datasets (the Datasaurus, plus 12 others) each have the same summary statistics (x/y mean, x/y standard deviation, and Pearson's correlation) to two decimal places, while being drastically different in appearance. This work describes the technique we developed to create this dataset, and others like it.

Fig 2. The Datasaurus Dozen. While different in appearance, each dataset has the same summary statistics (mean, standard deviation, and Pearson's correlation) to two decimal places.

The plots have identical values for popular metrics such as R-squared.



Chapter 10 Indices of Agreement

Several of these metrics are popular and do not relate to any important question.

You must use a metric that you understand, that your audience understands, and that relates to your research question.

$$E = 1 - \frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = 1 - \frac{\sum_{i=1}^{N} D_i{}^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = 1 - \frac{RMSD^2}{\text{Variance in } \mathbf{X}} \qquad \text{Equation 10.12}$$

$$E1 = 1 - \frac{\sum_{i=1}^{N}|X_i - Y_i|}{\sum_{i=1}^{N}|X_i - \bar{X}|} = 1 - \frac{\sum_{i=1}^{N}|D_i|}{\sum_{i=1}^{N}|X_i - \bar{X}|} \qquad \text{Equation 10.13}$$

$$dr = \begin{cases} 1 - \frac{\sum_{i=1}^{N}|D_i|}{2\sum_{i=1}^{N}|X_i - \bar{X}|} \text{ when } \sum_{i=1}^{N}|D_i| \le 2\sum_{i=1}^{N}|X_i - \bar{X}| \\ \frac{2\sum_{i=1}^{N}|X_i - \bar{X}|}{\sum_{i=1}^{N}|D_i|} - 1 \text{ when } \sum_{i=1}^{N}|D_i| > 2\sum_{i=1}^{N}|X_i - \bar{X}| \end{cases} \qquad \text{Equation 10.14}$$

$$M = \left(\frac{2}{\pi}\right) \text{ARCSIN}\left[1 - \frac{\sum_{i=1}^{N} D_i{}^2}{\sum_{i=1}^{N}[(X_i - \bar{X})^2 + (Y_i - \bar{Y})^2 + \bar{D}^2]}\right] \qquad \text{Equation 10.15}$$

$$\Re = 1 - \frac{N\sum_{i=1}^{N}|Y_i - X_i|}{\sum_{j=1}^{N}\sum_{i=1}^{N}|Y_j - X_i|} = 1 - \frac{\sum_{i=1}^{N}|D_i|}{\sum_{j=1}^{N}\sum_{i=1}^{N}|Y_j - X_i|/N} \qquad \text{Equation 10.16}$$

$$A = 1 - \frac{\sum_{i=1}^{N} D_i{}^2}{\sum_{i=1}^{N}[(2X_i - \bar{X} - \bar{Y})^2 + (2Y_i - \bar{X} - \bar{Y})^2]/2} \qquad \text{Equation 10.17}$$

$$AC = 1 - \frac{\sum_{i=1}^{N} D_i{}^2}{\sum_{i=1}^{N}[(|\bar{D}| + |X_i - \bar{X}|)(|\bar{D}| + |Y_i - \bar{Y}|)]} \qquad \text{Equation 10.18}$$

Report the unreliability in the Reference data.

Your reference data might be unreliable to the degree that "correct" and "error" make no sense.

Ground Truth in Classification Accuracy Assessment: Myth and Reality

Giles M. Foody

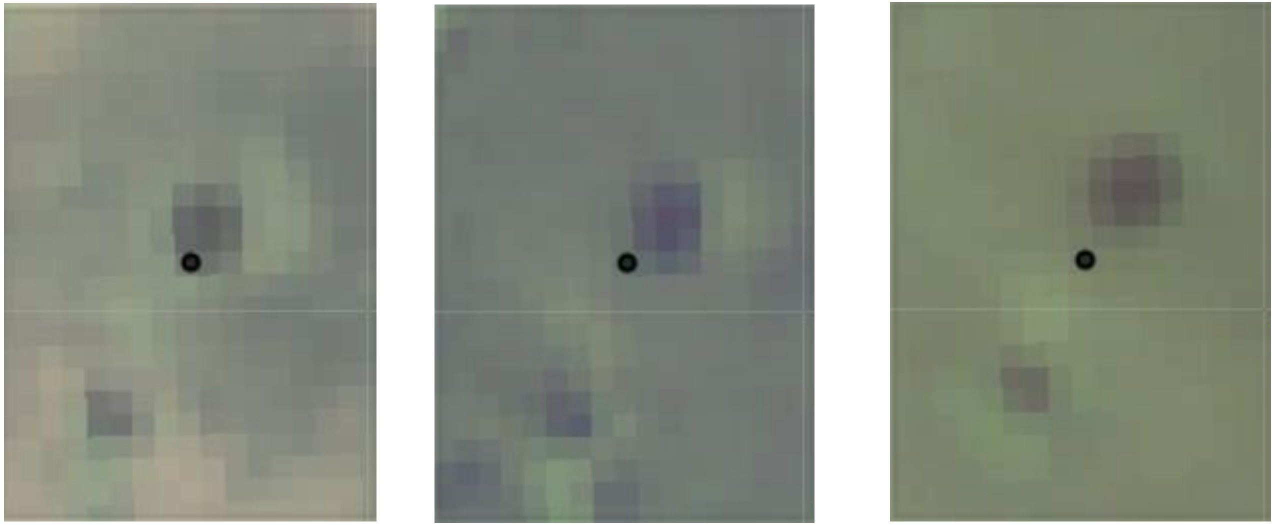
School of Geography, University of Nottingham, Nottingham NG7 2RD, UK; giles.foody@nottingham.ac.uk

Abstract: The ground reference dataset used in the assessment of classification accuracy is typically assumed implicitly to be perfect (i.e., 100% correct and representing ground truth). Rarely is this assumption valid, and errors in the ground dataset can cause the apparent accuracy of a classification to differ greatly from reality. The effect of variations in the quality in the ground dataset and of class abundance on accuracy assessment is explored. Using simulations of realistic scenarios encountered in remote sensing, it is shown that substantial bias can be introduced into a study through the use of an imperfect ground dataset. Specifically, estimates of accuracy on a per-class and overall basis, as well as of a derived variable, class areal extent, can be biased as a result of ground data error. The specific impacts of ground data error vary with the magnitude and nature of the errors, as well as the relative abundance of the classes. The community is urged to be wary of direct interpretation of accuracy assessments and to seek to address the problems that arise from the use of imperfect ground data.

https://doi.org/10.3390/geomatics4010005

Is there change of water at this sample point?

Time 1 | Time 2 | Time 3



Aiyin Zhang leads a team of students at Clark University.

NATIONAL SCIENCE FOUNDATION LTER NETWORK LONG TERM ECOLOGICAL RESEARCH

The images are inconsistently georegistered.
Various interpreters give different assessments.
Interpreters are uncertain, which means the reference data are unreliable.

Zhang, Muda, Domingues, and Pontius. (2024). Association of American Geographers.

Our profession’s leaders are informing our community.

ELSEVIER

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Remote Sensing of Environment

Review

Good practices for estimating area and assessing accuracy of land change

CrossMark

Pontus Olofsson [a,*], Giles M. Foody [b], Martin Herold [c], Stephen V. Stehman [d], Curtis E. Woodcock [a], Michael A. Wulder [e]

https://www.sciencedirect.com/science/article/abs/pii/S0034425714000704?via%3Dihub

Brave scientists report user's and producer's accuracies of less than 20% for land **change** at fine resolutions.

ELSEVIER

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Remote Sensing of Environment

Check for updates

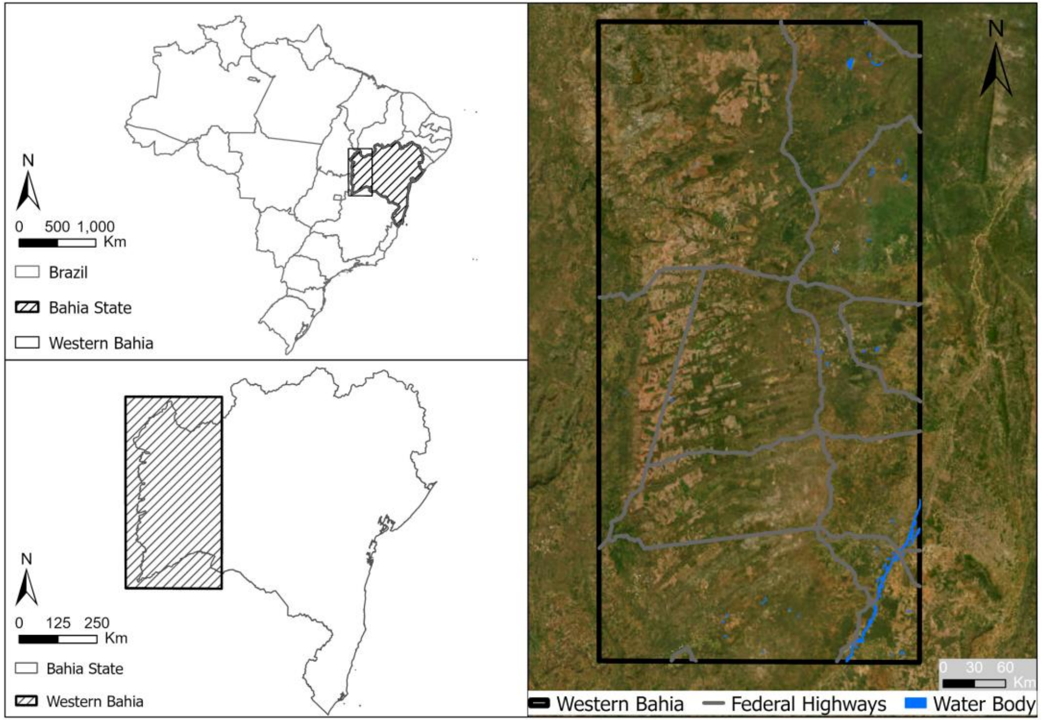
Validation of the U.S. Geological Survey's Land Change Monitoring, Assessment and Projection (LCMAP) Collection 1.0 annual land cover products 1985–2017

Stephen V. Stehman[a,*], Bruce W. Pengra[b], Josephine A. Horton[c], Danika F. Wellington[b]

[a] *College of Environmental Science and Forestry, State University of New York, Syracuse, NY 13210, USA*
[b] *KBR, contractor to the U.S. Geological Survey, Earth Resources Observation and Science (EROS) Center, Sioux Falls, SD 57198, USA*
[c] *Innovate! Inc., contractor to the U.S. Geological Survey EROS Center, Sioux Falls, SD 57198, USA*

https://www.sciencedirect.com/science/article/abs/pii/S0034425721003667?via%3Dihub

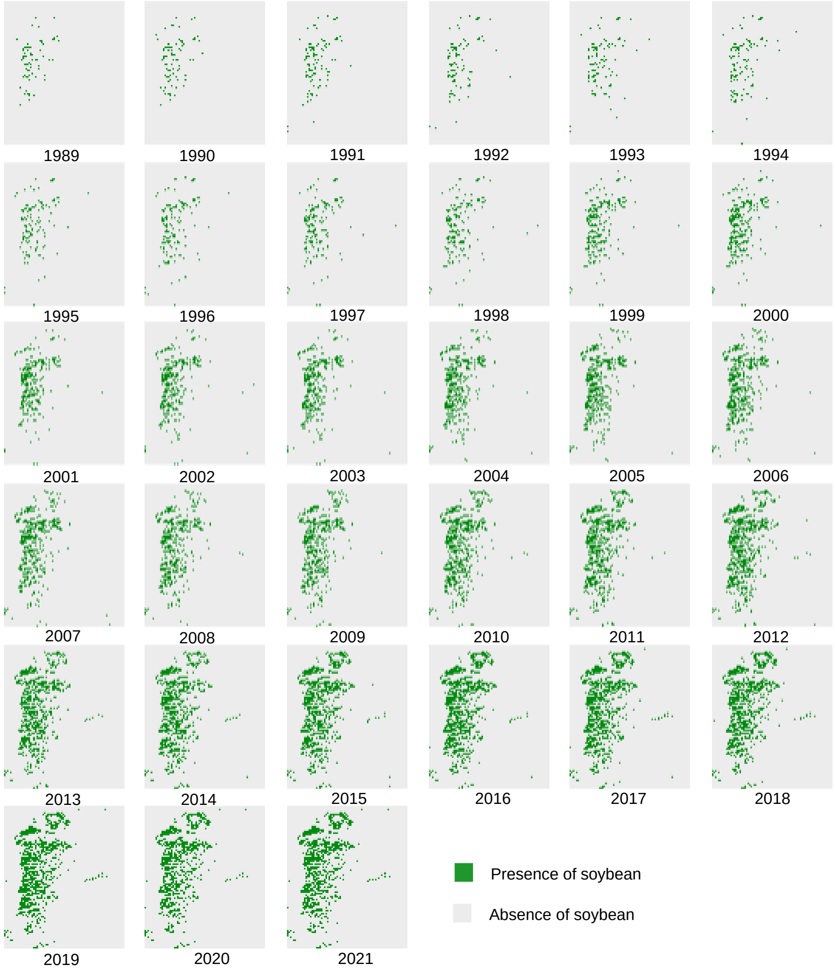
Western Bahia Brazil is a hotspot for soybean cultivation. Do the data make intuitive sense?



MAPBIOMAS

Pontius Jr, Robert Gilmore, Thomas Bilintoh, Gustavo de L. T. Oliveira, Julia Z. Shimbo. 2023. TRAJECTORIES OF LOSSES AND GAINS OF SOYBEAN CULTIVATION DURING MULTIPLE TIME INTERVALS IN WESTERN BAHIA, BRAZIL. Space Week Nordeste. Fortaleza, Brazil.

Maps show soybean at 33 years.

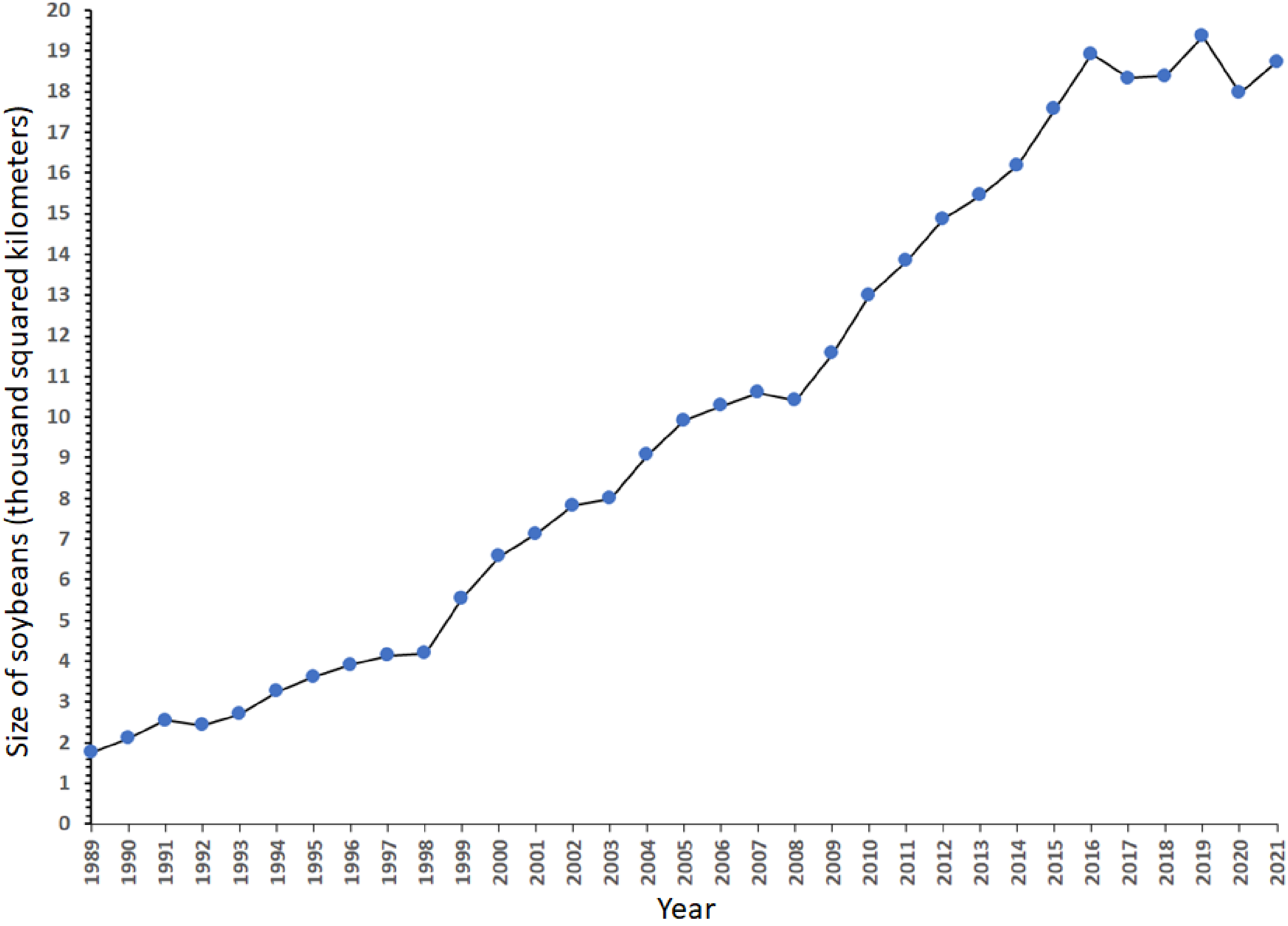


The extent has more than 200 million pixels. Each pixel has more than 8 billion possible combinations of presence or absence of soybean.

Reference data are too costly to collect.

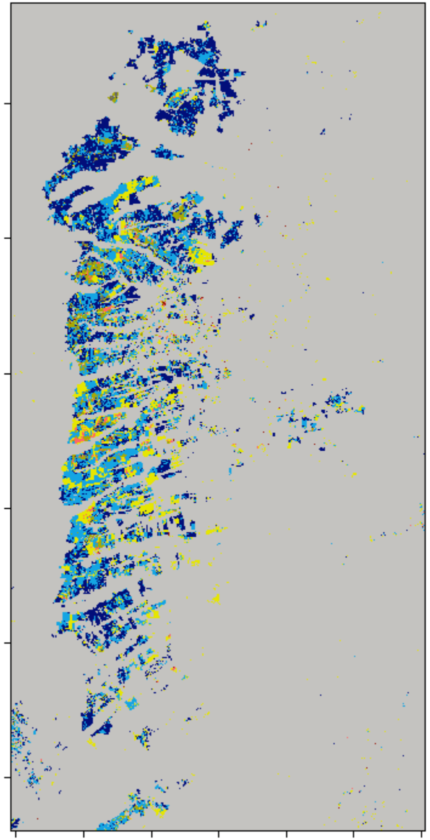
We must design a method to see whether the data make intuitive sense.

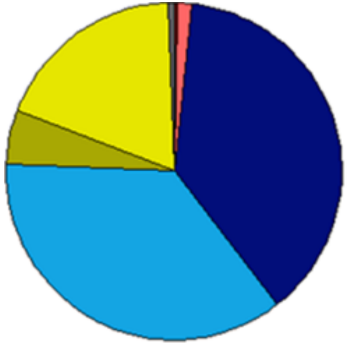
This popular format shows quantity, but fails to show allocation, alternation, or reliability.



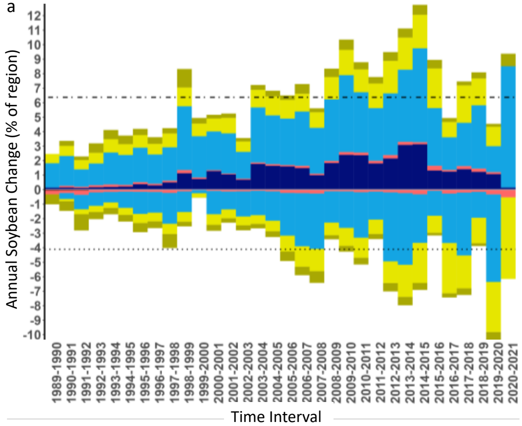
One map shows eight trajectories during 32 time intervals.

Visit the GitHub site of Thomas Bilintoh.





Most of the change is Alternation.



Loss Without Alternation
Loss With Alternation
Gain Without Alternation
Gain With Alternation
All Alternation Loss First
All Alternation Gain First
Stable Presence
Stable Absence

Get materials for free

Use free software packages at

https://cran.r-project.org/web/packages/diffeR/index.html

https://cran.r-project.org/web/packages/TOC/

https://lazygis.github.io/projects/TOCCurveGenerator

https://github.com/bilintoh/timeseriesTrajectories

Use PontiusMatrix42.xlsx at

http://www2.clarku.edu/~rpontius/

See videos at

https://www2.clarku.edu/faculty/rpontius/videos.html

Ali Santacruz
PhD 2014

Zhen Liu, M.A./GIS '21

Pontius' recommendations for Best Practices

1. Select a metric that addresses your research question, which is difficult.
2. Think in terms of quantity and allocation differences, which are concepts that popular metrics fail to distinguish.
3. Use the book Metrics That Make a Difference: How to Analyze Change and Error starting with the chapter *Commandments to Avoid Deadly Sins*.
4. Consider your motivations, which might conform to a flawed culture that reports accuracy without reporting the reference data's unreliability.
5. Get free materials at Pontius' website www.clarku.edu/~rpontius
6. Advise predoctoral colleagues to enter university programs, e.g. Clark University.
7. Discuss your problems openly to maximize learning.

We invited land-change modelers to submit:

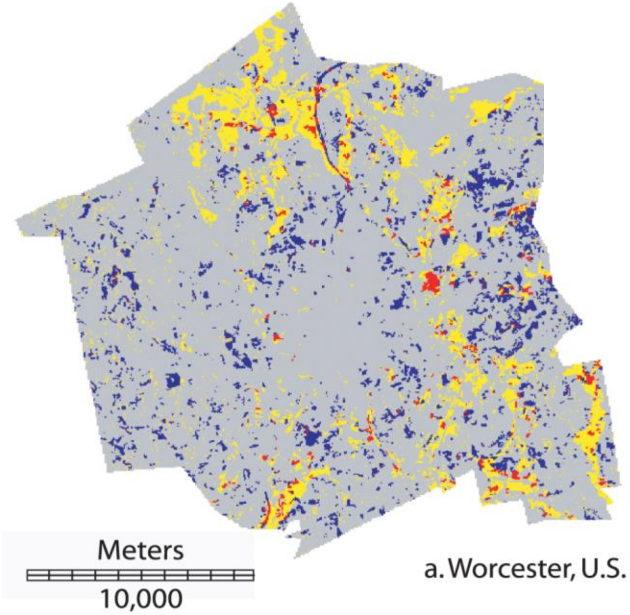
1. Reference Map of Time 1,
2. Reference Map of Time 2,
3. Prediction Map of Time 2,
4. Criterion to evaluate the maps.

We got some immediate interesting results:

1. Many scientists promised to send the maps.
2. Few of those scientists sent the maps.
3. Of the scientists who sent the maps, few sent any criterion.
4. Those who sent criterion usually sent percent correct between Reference and Prediction at time 2.

Pontius Jr et al. 2018. Lessons and Challenges in Land Change Modeling Derived from Synthesis of Cross-Case Comparisons. Chapter 8 in Martin Behnisch and Gotthard Meine (eds.) Trends in Spatial Analysis and Modelling. Geotechnologies and the Environment 19: 143-164. Springer International Publishing: Cham, Germany.

The Geomod Land Change Model Applied in the USA



There is more error than correctly predicted change.

Most of the error is due to predicting the wrong allocation by not more than 4 kilometers.

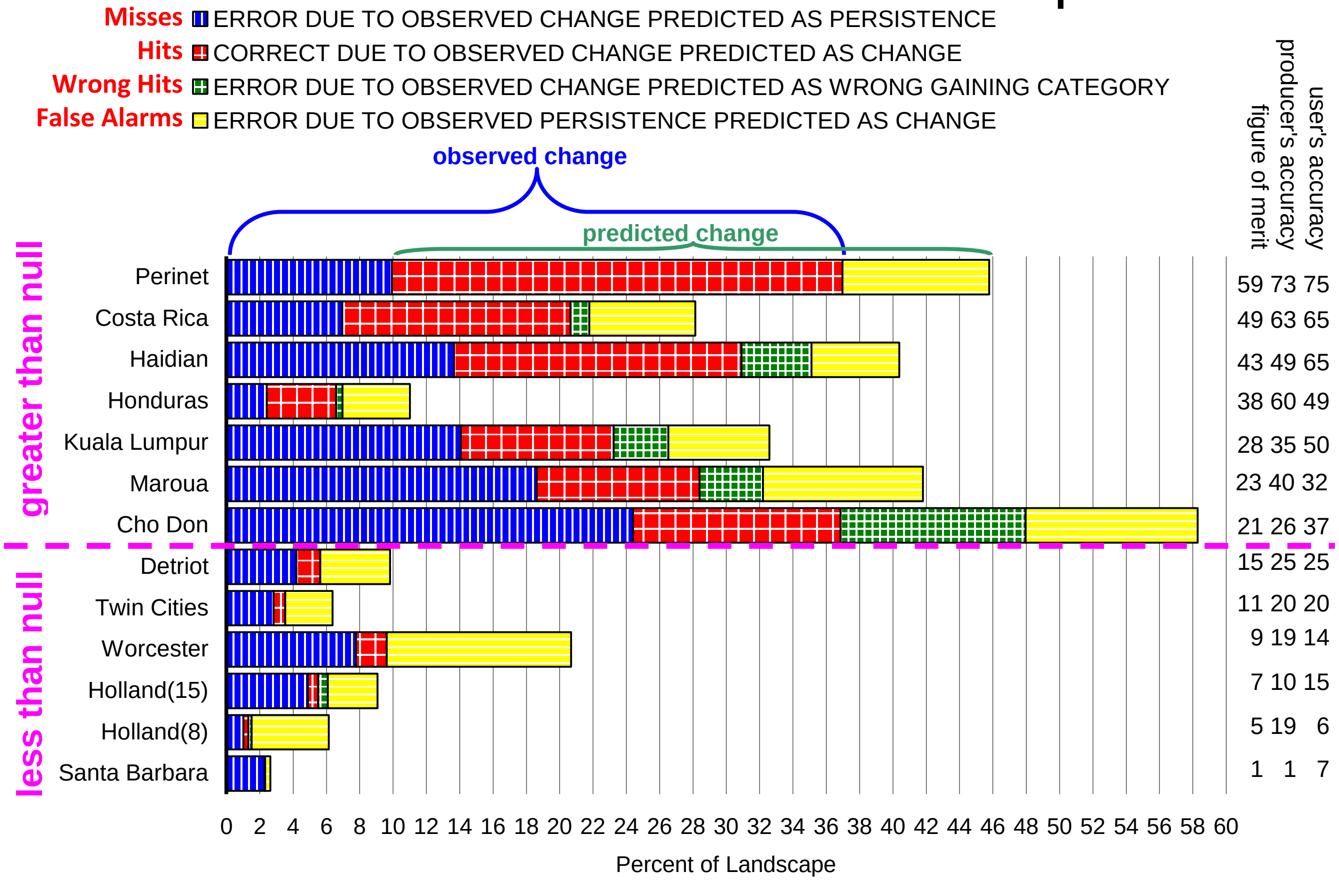
Misses	ERROR DUE TO OBSERVED CHANGE PREDICTED AS PERSISTENCE
Hits	CORRECT DUE TO OBSERVED CHANGE PREDICTED AS CHANGE
Wrong Hits	ERROR DUE TO OBSERVED CHANGE PREDICTED AS WRONG GAINING CATEGORY
False Alarms	ERROR DUE TO OBSERVED PERSISTENCE PREDICTED AS CHANGE
Correct Rejections	CORRECT DUE TO OBSERVED PERSISTENCE PREDICTED AS PERSISTENCE
	NOT CANDIDATE FOR TRANSITION
	OUT OF STUDY AREA

Thirteen applications shows that 12 had more error than hits at the resolution of the data.

Pontius Jr et al. 2018. Lessons and Challenges in Land Change Modeling Derived from Synthesis of Cross-Case Comparisons. Chapter 8 in Martin Behnisch and Gotthard Meine (eds.) Trends in Spatial Analysis and Modelling. Geotechnologies and the Environment 19: 143-164. Springer International Publishing: Cham, Germany.

12 of 13 cases had more error than hits.

Results reflect the data format rather than the predictive algorithm



Response from non-modelers

“Your colleagues must hate you!”

Response from modelers

"Thank you for exposing this,
because now I can publish any results!"